Université Paris I, Panthéon - Sorbonne

Première année Master M.A.E.F. 2007 – 2008

Statistiques

Plan du cours

1. Quelques rappels de théorie de la mesure.
2. Quelques rappels sur les applications de la théorie de la mesure aux probabilités.
3. Estimation paramétrique.
4. Tests paramétriques.

Bibliographie

• Livres pour revoir les bases....

1. Baillargeon, B. Probabilités, statistiques et techniques de régression. SMG.
2. Bercu, B., Pamphile, P. et Azoulay, E. Probabilités Applications - Cours Exercices. Edisciences.
3. Dress, F. Probabilités et Statistique. Dunod.
4. Lecoutre, J.-P. Statistiques et Probabilités. Dunod.

Théorie de la mesure et applications aux probabilités

– Ansel et Ducel, Exercices corrigés en théorie de la mesure et de l'intégration, Ellipses.
– Barbe, P. et Ledoux, M., Probabilités, Belin.
– Dacunha-Castelle, D. et Duflo, M., Probabilités et Statistiques (I), Masson
– Jacod, J., Cours d'intégration, http://www.proba.jussieu.fr/pageperso/jacod.html.
– Jacod, J., Cours de Probabilités, http://www.proba.jussieu.fr/pageperso/jacod.html.
– Toulouse, P. Thèmes de probabilités et statistiques, Masson.

Statistiques inférentielles

– Dacunha-Castelle, D. et Duflo, M., Probabilités et Statistiques (I), Masson.
– Fourdrinier, D., Statistique inférentielle, Dunod.
– Lecoutre, J.-M. et Tassi, P., Statistique non paramétrique et robustesse, Economica.
– Milhaud, X., Statistique, Belin.
– Monfort, A., Cours de statistique mathématique, Economica.
– Saporta, G., Probabilités, analyse des données et statistiques. Technip.
– Tsybakov, A. Introduction à la statistique non-paramétrique. Collection Mathématiques et Applications, Springer.

Cours de STATISTIQUES1

1 Rappels sur la théorie de la mesure

Introduction

Il demeure des choses inconnues à partir des connaissances antérieures. Il s'agit de la notion de probabilité
- Qu'est-ce qu'un événement et l'ensemble de tous les événements?
- Que se passe-t-il pour des probabilités d'événements moins classiques (par exemple l'ensemble des décimaux)?
- Comment traiter une variable aléatoire qui est continue et discrète à la fois (par exemple le nombre de minutes passées à regarder la TV)?

1.1 Mesures

1.1.1 Tribus

Notation. – Ω est un ensemble (fini ou infini).
- $\mathcal{P}(\Omega)$ est l'ensemble de tous les sous-ensembles (parties) de Ω.

Rappel. Soit E un ensemble. E est dénombrable s'il existe une bijection entre $\mathbb{N}$ et E ou un sous-ensemble de $\mathbb{N}$. Par exemple, un ensemble fini, $\mathbb{Z}$, $\mathbb{Z}\times\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{Z}^n$ sont dénombrables. En revanche, $\mathbb{R}$ n'est pas dénombrable.

Définition. Soit $\mathcal{F}$ une famille de parties de Ω (donc $\mathcal{F}\subset\mathcal{P}(\Omega)$). On dit que $\mathcal{F}$ est une algèbre si:
- $\Omega\in\mathcal{F}$;
- lorsque $A\in\mathcal{F}$ alors $(\Omega\setminus A)\in\mathcal{F}$;
- pour tout $n\in\mathbb{N}^*$, lorsque $(A_1,\dots,A_n)\in\mathcal{F}^n$ alors $A_1\cup\dots\cup A_n\in\mathcal{F}$.

Définition. Soit $\mathcal{A}$ une famille de parties de Ω (donc $\mathcal{A}\subset\mathcal{P}(\Omega)$). On dit que $\mathcal{A}$ est une tribu (ou une σ-algèbre) sur Ω si:
- $\Omega\in\mathcal{A}$;
- lorsque $A\in\mathcal{A}$ alors $(\Omega\setminus A)\in\mathcal{A}$;
- pour $I\subset\mathbb{N}$, lorsque $(A_i)_{i\in I}\in\mathcal{A}^I$ alors $\bigcup_{i\in I}A_i\in\mathcal{A}$.

Exemple. – Cas du Pile ou Face.
- Cas où Ω est infini : $\Omega=\mathbb{N}$ par exemple.

Propriété. Avec les notations précédentes:
1. $\emptyset\in\mathcal{A}$;
2. si A et B sont dans la tribu $\mathcal{A}$, alors $A\cap B$ est dans $\mathcal{A}$;
3. si $\mathcal{A}_1$ et $\mathcal{A}_2$ sont deux tribus sur Ω, alors $\mathcal{A}_1\cap\mathcal{A}_2$ est une tribu sur Ω. Plus généralement, pour $I\subset\mathbb{N}$, si $(\mathcal{A}_i)_{i\in I}$ est une famille de tribus sur Ω, alors $\bigcap_{i\in I}\mathcal{A}_i$ est une tribu sur Ω;
4. si $\mathcal{A}_1$ et $\mathcal{A}_2$ sont deux tribus sur Ω, alors $\mathcal{A}_1\cup\mathcal{A}_2$ n'est pas forcément une tribu sur Ω. Par exemple, si $\Omega=\{0,1,2\}$, $\mathcal{T}_1=\{\emptyset,\{0\},\{1,2\},\Omega\}$ et $\mathcal{T}_2=\{\emptyset,\{0,1\},\{2\},\Omega\}$.

Définition. Si E est une famille de parties de Ω (donc $E\subset\mathcal{P}(\Omega)$), alors on appelle tribu engendrée par E, notée $\sigma(E)$, la tribu engendrée par l'intersection de toutes les tribus sur E (on peut faire la même chose avec des algèbres).

Remarque. La tribu engendrée est la "plus petite" tribu (au sens de l'inclusion) contenant la famille.

Rappel. – Un ensemble ou vert U dans un espace métrique X est tel que pour tout $x\in U$, il existe $r>0$ tel que $B(x,r)\subset U$.
- On dit qu'un ensemble dans un espace métrique X est fermé si son complémentaire dans X est ouvert.

Définition. Soit Ω un ensemble quelconque. On appelle tribu borélienne sur Ω, notée $\mathcal{B}(\Omega)$, la tribu engendrée par les ouverts de Ω. Un ensemble de $\mathcal{B}(\Omega)$ est appelé borélien.

Exemple. – Boréliens sur $\mathbb{R}$, sur $]0,1[$. Attention $\bigcup_{x\in\left]\frac{1}{4},\frac{1}{4}\right[}\left]x-\frac{1}{10};x+\frac{1}{10}\right[\in \mathcal{B}(]0,1[)$ car une réunion quelconque d'ouverts est ouverte.

– Boréliens sur $\mathbb{R}^2$.

1.1.2 Espace mesurable

Définition. Soit Ω un ensemble et $\mathcal{A}$ une tribu sur Ω. On dit que $(\Omega,\mathcal{A})$ est un espace mesurable.

Contrairement à ce qu'on s'intéressera aux probabilités, on dira que $(\Omega,\mathcal{A})$ est un espace probabilisable.

Proposition. Si $(\Omega_i,\mathcal{A}_i)_i$ sont n espaces mesurables, alors un ensemble de la forme $A_1\times\cdots\times A_n$ où chaque $A_i\in\mathcal{A}_i$. L'ensemble des ensembles de cette forme est une réunion finie d'ensembles $A_1\times\cdots\times A_n$ où chaque $A_i\in\mathcal{A}_i$. L'ensemble des ensembles de cette forme est une algèbre et on note $\mathcal{A}_1\otimes\cdots\otimes\mathcal{A}_n$ (on dit $\mathcal{A}_1$ tensoriel $\mathcal{A}_2$ ... tensoriel $\mathcal{A}_n$) la tribu sur Ω engendrée par ces ensembles élémentaires.

Exemple. Pavés de $\mathbb{R}^d$.

Définition. On appelle espace mesurable produit, des $(\Omega_i,\mathcal{A}_i)$, l'espace mesurable $\left(\prod_{i=1}^{n}\Omega_i,\ \bigotimes_{i=1}^{n}\mathcal{A}_i\right)$.

Exemple. Pile / Face 2 fois.

1.1.3 Définitions et Propriété d'une mesure

Définition. Soit $(\Omega,\mathcal{A})$ un espace mesurable. L'application $\mu:\mathcal{A}\to[0,+\infty]$ est une mesure si:

– $\mu(\emptyset)=0$.

– Pour tout $I\subset\mathbb{N}$ et pour $(A_i)_{i\in I}$ famille disjointe de $\mathcal{A}$ (telle que $A_i\cup A_j=\emptyset$ pour $i\neq j$), alors

$$\mu\left(\bigcup_{i\in I}A_i\right)=\sum_{i\in I}\mu(A_i)$$ (propriété dite de σ-additivité)

Définition. Avec les notations précédentes:

– Si $\mu(\Omega)<+\infty$, on dit que μ est finie.

– Si $\mu(\Omega)<M$ avec $M<+\infty$, on dit que μ est bornée.

– Si $\mu(\Omega)=1$, on dit que μ est une mesure de probabilité.

Exemple. Cas de $\Omega=\mathbb{R}$, de $\mathbb{N}$, ou $\mathbb{R}^2$.

Définition. Si $(\Omega,\mathcal{A})$ est un espace mesurable (respectivement probabilisable) alors $(\Omega,\mathcal{A},\mu)$ est un espace mesuré (resp. probabilisé) quand μ est une probabilité.

Remarque. Sur $(\Omega,\mathcal{A})$, on peut définir une infinité de mesures.

Propriété. Soit $(\Omega,\mathcal{A},\mu)$ un espace mesuré et $(A_i)_{i\in\mathbb{N}}$ une famille de $\mathcal{A}$.

1. Si $A_1\subset A_2$, alors $\mu(A_1)\leq\mu(A_2)$.
2. Si $\mu(A_1)<+\infty$ et $\mu(A_2)<+\infty$, alors $\mu(A_1\cup A_2)+\mu(A_1\cap A_2)=\mu(A_1)+\mu(A_2)$.
3. Pour tout $I\subset\mathbb{N}$, on a $\mu\left(\bigcup_{i\in I}A_i\right)\leq\sum_{i\in I}\mu(A_i)$.
4. Si $A_i\subset A_{i+1}$ pour tout $i\in\mathbb{N}$ (suite croissante en sens de l'inclusion), alors $(\mu(A_n))_{n\in\mathbb{N}}$ est une suite croissante convergente tel que

$$\mu\left(\bigcup_{i\in\mathbb{N}}A_i\right)=\lim_{i\to+\infty}\mu(A_i)$$ (même si cette limite est $+\infty$)

5. Si $A_{i+1}\subset A_i$ pour tout $i\in\mathbb{N}$ (suite décroissante en sens de l'inclusion) et $\mu(A_0)<+\infty$, alors

$(\mu(A_n))_{n\in\mathbb{N}}$ est une suite décroissante convergente tel que $\mu\left(\bigcap_{i\in\mathbb{N}}A_i\right)=\lim_{i\to+\infty}\mu(A_i)$.

Exemple. 1. Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré. On définit $\nu(A) = \mu(A \cap B)$ où $B \in \mathcal{A}$. ν mesure?

2. Si μ_1 et μ_2 mesures sur $(\Omega, \mathcal{A})$, $\mu_1 + \mu_2$ et $\alpha\mu$ sont-elles des mesures?

Définitions. Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré et $(A_i)_{i \in \mathbb{N}}$ une famille de $\mathcal{A}$.

1. On définit $\limsup_n (A_n) = \bigcap_{n \in \mathbb{N}} \bigcup_{m \geq n} A_m$ (intuitivement, $\limsup_n (A_n)$ est l'ensemble des $\omega \in \Omega$ tels que ω appartient à une infinité de A_n).
2. On définit $\liminf_n (A_n) = \bigcup_{n \in \mathbb{N}} \bigcap_{m \geq n} A_m$ (intuitivement, $\liminf_n (A_n)$ est l'ensemble des $\omega \in \Omega$ tels que ω appartient à tous sauf à un nombre fini d'entre eux).

Exemple. Cas des suites croissantes ou décroissantes d'ensembles.

Théorème (Théorème d'extension de Hahn - Carathéodory). Si Ω est un ensemble, $\mathcal{F}$ une algèbre sur Ω, et ν une application de $\mathcal{F}$ dans $[0, +\infty]$ additive (tel que $\nu(A \cup B) = \nu(A) + \nu(B)$ pour $A \cap B = \emptyset$), alors si $\mathcal{A}$ est la tribu engendrée par $\mathcal{F}$, il existe une mesure ν sur la tribu $\mathcal{A}$ qui coïncide avec ν sur $\mathcal{F}$ (c'est-à-dire que pour tout $F \in \mathcal{F}$, $\nu(F) = \nu(F)$). On dit que ν prolonge ν sur la tribu $\mathcal{A}$.

Exemple. Définition de la mesure de Lebesgue sur $\mathbb{R}^n$, $n \geq 1$, $\mathbb{R}$, $\mathbb{R}$

Définitions. Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré

1. Pour $A \in \mathcal{A}$, on dit que A est μ-négligeable si $\mu(A) = 0$.
2. Soit une propriété P dépendant des éléments ω de Ω. On dit que P est vraie μ- presque partout (ou presque sûrement sur un espace probabilisé) si l'ensemble des ω pour lequel le elle n'est pas vérifiée est μ - négligeable.

Exemple. – Mesure de Lebesgue sur $\mathbb{R}$, $\mathbb{N}$ ou $\mathbb{Q}$

– La propriété " la suite de fonctions $(f_n(x) = x^n)$ converge vers la fonction $f(x) = 0$" est vraie λ- presque partout sur $[0, 1]$.

– Soit $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ et soit F la fonction définie par $F(x) = \mu(]-\infty, x])$ pour $x \in \mathbb{R}$.

1.1.4 Fonctions mesurables

Rappel. Soit $f : E \to F$, où E et F sont 2 espaces quelconques.

– Pour $I \subset F$, on appelle ensemble image réciproque de I par f, l'ensemble $f^{-1}(I) = \{x \in E, f(x) \in I\}$.

– (f continue) $\Rightarrow$ (pour tout ouvert U de F alors $f^{-1}(U)$ est un ouvert de E).

Définition. Soit $f : E \to F$ et soit $\mathcal{I}$ une tribu sur F. On note $f^{-1}(\mathcal{I})$ l'ensemble de sous-ensembles de Ω tel que $f^{-1}(\mathcal{I}) = \{f^{-1}(I), I \in \mathcal{I}\}$.

Propriété. Soit $(\Omega', \mathcal{A}')$ un espace mesurable et soit $f : \Omega \to \Omega'$. Alors $f^{-1}(\mathcal{A}')$ est une tribu sur Ω, appelée tribu engendrée par f.

Définitions. Soit $(\Omega, \mathcal{A})$ et $(\Omega', \mathcal{A}')$ deux espaces mesurables. Une fonction $f : \Omega \to \Omega'$ est dite mesurable pour les tribus $\mathcal{A}$ et $\mathcal{A}'$ si et seulement si $f^{-1}(\mathcal{A}') \subset \mathcal{A}$ (donc si et seulement $\forall A' \in \mathcal{A}'$, alors $f^{-1}(A') \in \mathcal{A}$).

Exemple. – Fonction indicatrice.

– Combinaison linéaire de fonctions indicatrices.

Remarque. Dans le cas où $(\Omega, \mathcal{A})$ est un espace probabilisable, et si $f : \Omega \to \mathbb{R}$, alors si f est une fonction mesurable sur $\mathcal{A}$ et $\mathcal{B}(\mathbb{R})$, alors f est une variable aléatoire.

Exemple. Nombre de Piles dans un jeu de Pile/Face.

Remarque. Dans le cas où $(\Omega, \mathcal{A})$ est un espace mesurable, et si $f : (\Omega, \mathcal{B}(\Omega'))$, où Ω' est un espace métrique et $\mathcal{B}(\Omega')$ l'ensemble des boréliens de Ω', si f est une fonction mesurable sur $\mathcal{B}(\Omega')$, alors f est dite fonction borélienne.

Propositions. Soit $(\Omega, \mathcal{A})$ et $(\Omega', \mathcal{A}')$ deux espaces mesurables et $f : \Omega \to \Omega'$. Soit $\mathcal{F}$ une famille de sous-ensembles de Ω' tel que $\mathcal{F}(\mathcal{F}) = \mathcal{A}'$. Alors

1. $f^{-1}(F)$ en gendre la t ribu $f^{-1}(A)$.
2. (f m es u rable) $\Leftrightarrow$ $(f^{-1}(F) \subset A)$

C o ns´qu e n ce. – Si (Ω, A) et (Ω', A') son t deu x espaces mesu rables et si $f : \Omega \to \Omega'$ est mes u rable, alors t ou te applicat ion cont in ue de Ω dans Ω' est mes u rable.

– Pou r mont rer qu ’u ne fonction $f : \Omega \to \mathbb{R}$ est mes u rable, il suffit de montrer que la famil le d’ensemble $(\{\omega \in \Omega, f(\omega) \leq a\})_{a \in \mathbb{R}} \in A$.

Pro priété. – Soit f mesurable de (Ω, A) dan s (Ω', A') et g m es u rable de (Ω', A') dan s (Ω'', A''). Alors $g \circ f$ est mesu rable dan s (Ω, A) et (Ω', A').

– Soit f_1 mes u rable de (Ω, A) dan s (Ω_1, A_1) et f_2 mes u rable de (Ω, A) dan s (Ω_2, A_2). Alor s $h : \Omega \to \Omega_1 \times \Omega_2$ tel le que $h(\omega) = (f_1(\omega), f_2(\omega))$ est mesurable dan s (Ω, A) et $A_1 \otimes A_2$.

– Soit $(f_n)_{n \in \mathbb{N}}$ une su it e de fonct ions mes u rables de (Ω, A) dans $(\Omega', B(\Omega'))$, o `u Ω' est u n es pace m ét riqu e, et tel le qu ’il ex ist e un e fon ct ion f limit e simple de la fonc $\forall \omega \in \Omega, \lim_{n \to \infty} f_n(\omega) = f(\omega)$). Alors f est mesu rable dan s (Ω, A) et $B(\Omega')$.

Défini ti on. S oit f mesu rable de (Ω, A) dan s (Ω', A') et soit $\mu : A' \to [0, +\infty]$ tel le que pour tou t $A' \in A'$, on ait $\mu_f(A') = \mu(f^{-1}(A'))$. Alor s μ_f est u n e mes u re su r (Ω', A'), appelé e mesu re im age de μ par f.

Cas parti cul ie r. S i μ est u n e mes u re de probabilit é s t u n e variable al éatoire alors μ_X est la mes u re (loi) de probabilit é de la var iable al éatoire X.

1 . 1.5 Cas des fonct i ons r´el l es m es u rabl es

Pro priété. Soit f et g deu x fonct ions r´el les mesurables (de (Ω, A) dans $(\mathbb{R}, B(\mathbb{R}))$). Alors $\alpha.f$, $f + g$, $\min(f, g)$ et $\max(f, g)$ sont des fonct ions r´el les mesurables.

Pro priété. Soit $(f_n)_{n \in \mathbb{N}}$ une su it e de fonct ions r´el les mesurables. Alors $\inf_n(f)$ et $\sup(f_n)$ son t des fon c- tions r´el les mesurables.

Dé fini ti o n. Soit $f : \Omega \to \mathbb{R}$. Alors f est dit e ét ag é s’il existe une famil le d’ensembles disjoints $(A_i)_{1 \leq i \leq n}$ de Ω et une famil le de r´el s $(\alpha_i)_{1 \leq i \leq n}$ tel les que pour tou t $\omega \in \Omega$, on ait $f(\omega) = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}(\omega)$.

R e m arqu e. Si les A_i sont t ou s dans A st ribu su r Ω, alors f est A- m es u rable.

Th éo r`e m e. Tou te fonct ion r´el le mesurable `a valeurs dans $\mathbb{R}^+$ est limit e simple d’u ne su it e croissan te de fonct ions ét ag é s.

C o ns´qu e n ce. Soit f une fonct ion r´el le mesurable. Alors f est limite simple de fonct ions ét ag é s.

1.2 Int égration de Leb esgue

Da ns to ute la suite, on co ns id `e re (Ω, A, μ) un espa ce mes ur´

1 . 2.1 Int é g ral e de Leb es gue d’ une foncti on p os i ti ve

Dé fini ti o n. 1. Soit $f = \mathbb{1}_a$, o `u $A \in A$. Alor s:

$$\int f d\mu = \int_{\omega} f(\omega) d\mu(\omega) = \mu(A).$$

2. Soit $f = \mathbb{1}_a$, o `u $A \in A$ et soit $B \in A$. Alor s:

$$\int_B f d\mu = \int_B f(\omega) d\mu(\omega) = \mathbb{1}_B \mu(A)(\omega) f(\omega) d\mu(\omega) = \mu(A \cap B).$$

3. S oit f une fonct ion ét ag é e positive tel le que $f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$, o `u les $A_i \in A$ et $\alpha_i > 0$ et soit $B \in A$.

Alor s:

$$\int_B f d\mu = \int_B f(\omega) d\mu(\omega) = \mathbb{1}_B(\omega) f(\omega) d\mu(\omega) = \sum_{i=1}^{n} \alpha_i \mu(A_i \cap B).$$

Exemple. Fonction $\mathbb{1}_{\mathbb{Q}}$, fonctions en escalier,...

Définition. Soit f une fonction $\mathcal{A}$- mesurable positive et soit $B \in \mathcal{A}$. Alors l'intégrale de Lebesgue de f par rapport à μ sur B est:

$$\int_B f\,d\mu = \int \mathbb{1}_B(\omega) f(\omega)\,d\mu(\omega) = \sup \int_B g\,d\mu, \text{ pour } g \text{ étagée positive telle que } g \leq f.$$

Propriété. Soit f une fonction $\mathcal{A}$- mesurable positive et soit $A \in \mathcal{A}$, $B \in \mathcal{A}$. Alors:

1. Pour $c \geq 0$, $\int_B cf\,d\mu = c\int_B f\,d\mu$.
2. Si $A \subset B$, alors $\int_A f\,d\mu \leq \int_B f\,d\mu$.
3. Si g est une fonction $\mathcal{A}$-mesurable positive telle que $0 \leq f \leq g$ alors $0 \leq \int_B f\,d\mu \leq \int_B g\,d\mu$.
4. Si $\mu(B) = 0$ alors $\int_B f\,d\mu = 0$.

Théorème (Théorème de convergence monotone (Beppo Levi)). Si $(f_n)_n$ est une suite croissante de fonctions mesurables positives convergeant simplement vers f sur Ω, alors:

$$\lim_{n\to\infty} \int f_n\,d\mu = \int f\,d\mu = \lim_{n\to\infty} \int f_n\,d\mu.$$

Conséquence. Pour les séries de fonctions mesurables positives, on peut toujours appliquer le théorème de convergence monotone et donc inverser la somme et l'intégrale.

Lemme (Lemme de Fatou). Soit $(f_n)_n$ est une suite de fonctions mesurables positives alors:

$$\liminf_{n\to\infty} \int f_n\,d\mu \leq \liminf_{n\to\infty} \int f_n\,d\mu.$$

Exemple. Appliquer Fatou à (f_n) tel que $f_{2n} = \mathbb{1}_A$ et $f_{2n+1} = \mathbb{1}_B$.

1.2.2 Intégrale de Lebesgue d'une fonction de signe quelconque et propriétés

Définition. Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré, $B \in \mathcal{A}$ et soit f une fonction $\mathcal{A}$- mesurable à valeurs réelles, telle que $f = f^+ - f^-$ avec $f^+ = \max(f, 0)$ et $f^- = \max(-f, 0)$. On dit que f est μ- intégrable sur B si $\int_B |f|\,d\mu < +\infty$. On a alors

$$\int_B f\,d\mu = \int_B f^+\,d\mu - \int_B f^-\,d\mu.$$

Notation. Lorsque f est μ- intégrable sur B, soit $\int |f|\,d\mu < +\infty$, on note $f \in L^1(\Omega, \mathcal{A}, \mu)$ (on dit que f est L^1).

Exemple. Intégrale de Riemann et intégrale de Lebesgue.
Cas de la masse de Dirac.

Propriété. On suppose que $f \in L^1(\Omega, \mathcal{A}, \mu)$. Alors:

1. $\int (\alpha f + \beta g)\,d\mu = \alpha \int f\,d\mu + \beta \int g\,d\mu$ pour $(\alpha, \beta) \in \mathbb{R}^2$.
2. Si $f \leq g$ alors $\int f\,d\mu \leq \int g\,d\mu$.

Théorème (Théorème de convergence dominée de Le Besgue). Soit $(f_n)_n$ est une suite de fonctions de $L^1(\Omega, \mathcal{A}, \mu)$ telles que pour tout $n \in \mathbb{N}$, $|f_n| \leq g$ avec $g \in L^1(\Omega, \mathcal{A}, \mu)$. Si on suppose que (f_n) converge simplement vers f sur Ω alors:

$$\lim_{n\to\infty} \int f_n\,d\mu = \int f\,d\mu.$$

Extension du Théorème de Lebesgue s'appliquant dans le cas où (f_n) converge presque partout vers f.

Exemple. Convergence d'une intégrale dépendant d'un paramètre : par exemple $\int_0^\infty \frac{f(x)}{1+x^n}dx$.

Théorème (Inégalité de Jensen). Soit (Ω, A, IP) un espace probabilisé, soit $\varphi : IR \to IR$ une fonction convexe et soit $f : \Omega \to IR$ mesurable tel le que $\varphi(f)$ soit une fonction intégrable par rapport à P. Alors:

$$\varphi\left(\int f dIP\right) \leq \int \varphi(f)\, dIP.$$

Exemple. Soit X une v.a. sur (Ω, IP). Alors $\varphi(IE X) \leq IE(\varphi(X))$.

1.2.3 Mesures induites et densité

Théorème (Théorème du Transport). Soit f une fonction mesurable de (Ω, A, μ) dans (Ω', A') tel le que μ soit la mesure induite par f ($\mu_f(A') = \mu(f^{-1}(A'))$ pour $A' \in A'$) et soit φ une fonction mesurable de (Ω', A') dans $(IR, B(IR))$. Alors, si $\varphi \in L^1(\Omega, A, \mu)$,

$$\int_{\Omega'} \varphi d\mu_f = \int_\Omega \varphi_0 f \, d\mu.$$

Définition. Soit μ et ν deux mesures sur (Ω, A). On dit que μ domine ν (ou ν est dominée par μ) et que ν est absolument continue par rapport à μ lorsque $\forall A \in A$, $\mu(A) = 0 \Rightarrow \nu(A) = 0$.

Propriété. Soit (Ω, A, μ) un espace mesuré et f une fonction définie sur (Ω, A) mesurable et positive. On suppose que pour $A \in A$, $\nu(A) = \int_A f \, d\mu$. Alors, ν est une mesure sur (Ω, A), dominée par μ. De plus, pour toute fonction g définie sur (Ω, A), mesurable et positive,

$$\int g d\nu = \int g.f \, d\mu.$$

Enfin, g est ν intégrable si et seulement si $g.f$ est μ intégrable.

Définition. On dit que μ mesure sur (Ω, A) est σ-finie lorsqu'il existe une famille $(A_i)_{i \in I}$, avec I dénombrable, d'ensembles A tel le que $A_i = \Omega$ et $\mu(A_i) < +\infty$ pour tout $i \in I$.

Théorème (Théorème de Radon-Nikodym). On suppose que μ et ν sont deux mesures σ-finies sur (Ω, A) tel les que μ domine ν. Alors, il existe une fonction définie sur (Ω, A) mesurable et positive, appelée densité de ν par rapport à μ, tel le que pour tout $A \in A$, $\nu(A) = \int_A f \, d\mu$.

Théorème (Théorème de Fubini). Soit $\Omega = \Omega_1 \times \Omega_2$, $A = A_1 \otimes A_2$ et $\mu = \mu_1 \otimes \mu_2$ (mesures σ finies), où (Ω_1, A_1, μ_1) et (Ω_2, A_2, μ_2) sont des espaces mesurés. Soit une fonction $f : \Omega \to IR$, A- mesurable et μ - intégrable. alors:

$$\int_\Omega f d\mu = \int_{\Omega_1}\left(\int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2(\omega_2)\right) d\mu_1(\omega_1) = \int_{\Omega_2}\left(\int_{\Omega_1} f(\omega_1, \omega_2) d\mu_1(\omega_1)\right) d\mu_2(\omega_2).$$

1.2.4 Espaces L^p

Définition. Soit (Ω, A, μ) un espace mesuré. On appelle l'espace $L^p(\Omega, A, \mu)$, où $p > 0$, l'ensemble des fonctions $f : \Omega \to IR$, mesurables et tel les que $\int |f|^p d\mu < +\infty$.

Définition. Pour $f \in L^p(\Omega, A, \mu)$, où $p > 0$, on note $\|f\|_p = \left(\int |f|^p d\mu\right)^{1/p}$.

Propriété (Inégalité de Hölder). Soit $p > 1$ et $q > 1$ tels que $\frac{1}{p} + \frac{1}{q} = 1$, et soit $f \in L^p(\Omega, A, \mu)$ et $g \in L^q(\Omega, A, \mu)$. Alors, $fg \in L^1(\Omega, A, \mu)$ et

$$\|fg\|_1 \leq \|f\|_p . \|g\|_q .$$

Propriété (Inégalité de Minkowski). Soit $p \geq 1$ et soit f et $g \in L^p(\Omega, \mathcal{A}, \mu)$. Alors, $f+g \in L^p(\Omega, \mathcal{A}, \mu)$ et

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p .$$

Remarque. Pour $p \geq 1$, $\|\cdot\|_p$ définie ainsi sur un espace semi-normé $L^p(\Omega, \mathcal{A}, \mu)$. Pour obtenir une norme, il faut se placer dans l'espace $L^p(\Omega, \mathcal{A}, \mu)$ obtenu en "quotientant" $L^p(\Omega, \mathcal{A}, \mu)$ par la relation d'équivalence $f = g$ μ-presque partout (c'est-à-dire que dans $L^p(\Omega, \mathcal{A}, \mu)$ on dira que $f = g$ lorsque $f = g$ μ-presque partout).

Définition. Pour f et $g \in L^2(\Omega, \mathcal{A}, \mu)$, on définit le produit scalaire $\langle f, g \rangle = \int fg \, d\mu$. On munit ainsi $L^2(\Omega, \mathcal{A}, \mu)$ d'une structure d'espace de Hilbert. On dira que f est orthogonale à g lorsque $\langle f, g \rangle = 0$.

Conséquence. Si A est un sous-espace vectoriel fermé de $L^2(\Omega, \mathcal{A}, \mu)$ (par exemple un sous-espace de dimension finie), alors pour tout $f \in L^2(\Omega, \mathcal{A}, \mu)$, il existe un unique projeté orthogonal de f sur A, noté f_A, qui vérifie $\|f - f_A\|_2 = \operatorname{Arg\,inf}_{g \in A} \|g - f\|_2$.

2 Applications de la théorie de la mesure et de l'intégration en Probabilités

2.1 Espérance de variables aléatoires

Définition. Soit X une variable aléatoire sur $(\Omega, I\!P)$ un espace probabilisé. Alors si $X \in L^1(\Omega, \mathcal{A}, I\!P)$, on définit l'espérance de X par le nombre $I\!E\, X = \int X \, dI\!P$. Plus généralement, si $\varphi : I\!R \to I\!R$ est borélien et si $\varphi(X) \in L^1(\Omega, \mathcal{A}, I\!P)$, on définit l'espérance de $\varphi(X)$ par $I\!E\, \varphi(X) = \int \varphi(X) \, dI\!P$.

Propriété. Si X est une variable aléatoire sur $(\Omega, I\!P)$, si $\varphi : I\!R \to I\!R$ est borélienne tel le que $\varphi(X) \in L^1(\Omega, \mathcal{A}, I\!P)$, et si $I\!P_X$ est la mesure de probabilité X alors:

$$I\!E\, \varphi(X) = \int_{I\!R} \varphi(x) \, dI\!P_X(x).$$

Conséquence. – Si $I\!P_X$ est absolument continue par rapport à la mesure de Lebesgue (donc X est une v.a. dite absolument continue), de densité f_X, alors $I\!E\, \varphi(X) = \int_{I\!R} \varphi(x) f_X(x) dx$.

– Si $I\!P_X$ est absolument continue par rapport à la mesure de comptage sur $I\!N$ (donc X est une v.a. discrète), de densité, alors $I\!E\, \varphi(X) = \sum_{k=0}^{\infty} p_X(k)\, \varphi(k)$.

Propriété.

1. Soit X et Y des variables aléatoires tel les que X et $Y \in L^1(\Omega, \mathcal{A}, I\!P)$. Alors pour tout $(a, b) \in I\!R^2$, $aX + bY \in L^1(\Omega, \mathcal{A}, I\!P)$ et
$$I\!E(aX + bY) = aI\!EX + b\, I\!EY.$$
2. Soit X une variable aléatoire sur $(\Omega, I\!P)$, et soit $A \in \mathcal{A}$. Alors $I\!E(\mathbb{1}_A(X)) = I\!P(X \in A)$.
3. Soit X et Y des variables aléatoires tel les que $X \in L^p(\Omega, \mathcal{A}, I\!P)$ et $Y \in L^q(\Omega, \mathcal{A}, I\!P)$ avec $\frac{1}{p} + \frac{1}{q} = 1$ et $p > 1$, $q > 1$. Alors $XY \in L^1(\Omega, \mathcal{A}, I\!P)$ et
$$I\!E|X.Y| \leq (I\!E|X|^p)^{1/p} (I\!E|Y|^q)^{1/q}.$$
4. Soit X et Y des variables aléatoires tel les que X et $Y \in L^p(\Omega, \mathcal{A}, I\!P)$, avec $p \geq 1$. Alors $X + Y \in L^p(\Omega, \mathcal{A}, I\!P)$ et
$$(I\!E|X+Y|^p)^{1/p} \leq (I\!E|X|^p)^{1/p} + (I\!E|Y|^p)^{1/p}.$$
5. Soit X une variable aléatoire tel le que $X \in L^p(\Omega, \mathcal{A}, I\!P)$ pour $p > 0$. Alors pour tout $0 < r \leq p$, $X \in L^r(\Omega, \mathcal{A}, I\!P)$ et
$$(I\!E|X|^r)^{1/r} \leq (I\!E|X|^p)^{1/p}.$$

6. Si X est une variable aléatoire sur (Ω, A, IP), si $\varphi : IR \to IR$ est une fonction convexe telle que X et $\varphi(X) \in IL^1(\Omega, A, IP)$, alors
$$IE(\varphi(X)) \geq \varphi(IE\, X).$$

Définition. Pour X et Y des variables aléatoires telles que X et $Y \in IL^2(\Omega, A, IP)$, on définit la covariance de X et Y par
$$cov(X,Y) = IE\,[(X - IE\, X)(Y - IE\, Y)];$$
On appelle la variance de X, $var(X) = cov(X, X) = IE\,(X - IEX)^2 = IE\,(X^2) - (IEX)^2$.

Propriété. Sur $IL^2(\Omega, A, IP)$, $cov(.,.)$ définit un produit scalaire. De plus
$$|cov(X, Y)|^2 \leq var(X).var(Y).$$

2.2 Fonction de répartition et quantiles d'une loi de probabilité

Il y a une correspondance bijective entre la connaissance de $F_X = IP_X(]-\infty, x])$. La fonction de répartition permet également de définir les quantiles qui sont essentiels à la construction d'intervalles de confiance et de test.

Soit $\alpha \in [0, 1]$. Des propriétés de la fonction de répartition, on en déduit qu'il existe $x_\alpha \in IR$, tel que:
$$\lim_{x \to x_\alpha} F_X(x) \leq \alpha \leq F_X(x_\alpha). \qquad (1)$$

Soit $I_\alpha = \{x_\alpha \in IR \text{ tel que } x_\alpha \text{ vérifie (1)}\}$. On appelle quantile (ou fractile, ou percentile en anglais) d'ordre α de la loi IP_X, noté q_α, le milieu de l'intervalle I_α. Il est évidemment, lorsque X admet une distribution absolument continue par rapport à la mesure de Lebesgue, égal à $F_X^{-1}(\alpha)$, où F_X^{-1} désigne la fonction inverse de F_X.

Deux cas particuliers sont à connaître:

1 / pour $\alpha = 0.5$, $q_{0.5}$ est appelé médiane de IP_X;

2 / pour $\alpha = 0.25$ et $\alpha = 0.75$ (respectivement), $q_{0.25}$ et $q_{0.75}$ sont appelés premier et troisième quartile (respectivement) de IP_X;

3 / pour $\alpha = 0.1, \ldots, 0.9$, on parle de déciles de IP_X.

2.3 Principales lois de probabilité

Loi uniforme discrète:

C'est la loi de probabilité associée à la valeur de $\{x_1, \ldots, x_n\}$ telle que
$$IP(X = x_i) = \frac{1}{n}.$$
On a alors : $IE\,X = \frac{1}{n}(x_1 + \ldots + x_n)$ et $var(X) = \frac{1}{n}(x_1^2 + \ldots + x_n^2) - (IE\,X)^2$.

Loi de Bernoulli:

C'est la loi de probabilité notée $B(p)$ à valeur dans $\{0, 1\}$ telle que
$$IP(X = 1) = p \quad \text{et} \quad IP(X = 0) = 1 - p.$$
On a alors : $IEX = p$ et $var(X) = p(1-p)$.

Loi binomiale:

C'est la loi de probabilité notée $B(n, p)$ à valeur dans $\{0, 1, \ldots, n\}$ telle que
$$IP(X = k) = C_n^k \cdot p^k \cdot (1-p)^{n-k} \quad \text{pour } k \in \{0, 1, \ldots, n\}$$

On a alors : $X = X_1 + \cdots + X_n$, où (X_i) est une suite de v.a. i.i.d. de Bernoulli de paramètre p, $\mathbb{E}X = np$ et $\mathrm{var}(X) = np(1-p)$.

Loi de Poisson:

C'est la loi de probabilité sur $\mathbb{N}$ notée $\mathcal{P}(\theta)$ à valeurs dans $\mathbb{N}$ telle que

$$\mathbb{P}(X = k) = \frac{\theta^k}{k!} \cdot e^{-\theta} \quad \text{pour } k \in \mathbb{N}.$$

On a alors $\mathbb{E}X = \theta$ et $\mathrm{var}(X) = \theta$.

Loi uniforme sur $[a, b]$

Cette loi est généralement notée $\mathcal{U}([a, b])$, où $-\infty < a < b < \infty$. C'est la loi de probabilité à valeurs dans $[a, b]$ de densité par rapport à la mesure de Lebesgue:

$$f_X(x) = \frac{1}{b-a} \cdot \mathbb{1}_{x \in [a,b]}$$

On a alors $\mathbb{E}X = \frac{b+a}{2}$ et $\mathrm{var}(X) = \frac{(b-a)^2}{12}$.

Loi Gamma:

Cette loi est généralement notée $\gamma(p, \theta)$, où $p > 0$ et $\theta > 0$. C'est la loi de probabilité à valeurs dans $\mathbb{R}_+$ de densité par rapport à la mesure de Lebesgue:

$$f_X(x) = \frac{\theta^p}{\Gamma(p)} \cdot e^{-\theta \cdot x} \cdot x^{p-1} \cdot \mathbb{1}_{x \in \mathbb{R}_+}.$$

On a alors $\mathbb{E}X = \frac{p}{\theta}$ et $\mathrm{var}(X) = \frac{p}{\theta^2}$.
Si $X \sim \gamma(p, \theta)$ et $Y \sim \gamma(q, \theta)$ avec X et Y indépendantes et $p > 0$ et $q > 0$, alors $X + Y \sim \gamma(p + q, \theta)$.
Pour $p = 1$, la loi $\gamma(p, \theta)$ est la loi exponentielle $\mathcal{E}(\theta)$.

Loi Bêta:

Cette loi est généralement notée $\beta(p, q)$, où $p > 0$ et $q > 0$. C'est la loi de probabilité à valeurs dans $[0, 1]$ de densité par rapport à la mesure de Lebesgue:

$$f_X(x) = \frac{x^p(1-x)^{q-1}}{B(p,q)} \cdot x^{p-1} \cdot \mathbb{1}_{x \in [0,1]} \text{ où } B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

On a alors $\mathbb{E}X = \frac{B(p+1,q)}{B(p,q)}$ et $\mathrm{var}(X) = \frac{p \cdot q}{(p+q)^2(p+q+1)}$.
Si $X \sim \gamma(p, \theta)$ et $Y \sim \gamma(q, \theta)$ avec X et Y indépendantes et $p > 0$ et $q > 0$, alors $\frac{X}{X+Y} \sim \beta(p, q)$.
Pour $p = 1$, la loi $\gamma(p, \theta)$ est la loi exponentielle $\mathcal{E}(\theta)$.

Loi normale (ou gaussienne) centrée réduite:

Cette loi est généralement notée $\mathcal{N}(0, 1)$. C'est la loi de probabilité à valeurs dans $\mathbb{R}$ de densité par rapport à la mesure de Lebesgue:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

On a :

$$\mathbb{E}(X) = 0 \text{ et } \mathrm{var}(X) = 1.$$

Loi normale (ou gaussienne) de moyenne m et de variance σ^2:

Si Z suit la loi $N(0,1)$, $X = m + \sigma Z$ suit par définition la loi $N(m, \sigma^2)$, loi normale d'espérance m et de variance σ^2. La densité de X est donnée par:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

La figure A.1. représente la densité de la loi normale centrée réduite et celle d'une loi normale non centrée et non réduite. A partir de la loi gaussienne, on peut définir les lois suivantes.

Loi du χ^2 à n degrés de liberté:

Soit $X_1, \cdots, X_n$, n variables aléatoires indépendantes de $N(0,1)$, alors

$$S = X_1^2 + \cdots + X_n^2$$

suit une loi du χ^2 à n degrés de liberté, loi notée $\chi^2(n)$. Cette loi est à valeurs dans $\mathbb{R}_+$, d'espérance n et de variance $2n$. C'est aussi la loi Gamma $\gamma(n/2, 1/2)$, c'est-à-dire $\chi^2(n)$ admet pour densité par rapport à la mesure de Lebesgue:

$$f_X(x) = \frac{1}{2^{n/2} \cdot \Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{x}{2}\right) \cdot \mathrm{II}_{\{x \geq 0\}},$$

où la fonction Gamma est telle que $\Gamma(a) = \int_0^\infty x^{a-1} \cdot e^{-x}$ pour $a \geq 0$. Enfin, si X suit une loi $\chi^2(n)$, par définition on dira que $Y = \sigma^2 X$ suit une loi $\sigma^2 \cdot \chi^2(n)$. La figure A.2. exhibe trois densités différentes de densité de loi du χ^2. Loi de Student à n degrés de liberté:

La loi de Student à n degrés de liberté, notée $T(n)$, est la loi du quotient

$$T = \frac{N}{\sqrt{S/n}}$$

où N suit une loi $N(0,1)$ et S suit une loi $\chi^2(n)$, N et S étant deux variables aléatoires indépendantes. Il est également possible de déterminer la densité d'une telle loi par rapport à la mesure de Lebesgue, à savoir,

$$f_X(x) = \frac{1}{\sqrt{n} \cdot B(1/2, n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2},$$

où la fonction Beta est telle que $B(a,b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}$ pour $a > 0$ et $b > 0$. La figure A.3. illustre deux exemples de cette densité, qu'il convient de comparer également avec la densité de la loi normale centrée réduite. Remarque : Par la loi des grands nombres, plus n est grand, plus S est proche de sa moyenne, qui vaut 1. Le dénominateur est donc proche de 1. Il s'ensuit que la loi $T(n)$ est d'autant plus proche d'une loi normale que n est grand.

Un des principaux intérêts de la loi de Student est dû dans le fait que, si $X_1, \cdots, X_n$ sont n variables aléatoires indépendantes de $N(m, \sigma^2)$, si on considère la moyenne et la variance empiriques:

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \quad \text{et} \quad \hat{\sigma}_n^2 = \frac{1}{n-1}\left((X_1 - \bar{X}_n)^2 + \cdots + (X_d - \bar{X}_n)^2\right),$$

alors

$$T = \frac{\sqrt{n} \cdot (\bar{X}_n - m)}{\hat{\sigma}_n}$$

suit une loi de Student à $n-1$ degrés de liberté.

Loi de Fisher à n_1 et n_2 degrés de liberté:

Soit S_1 et S_2 deux variables aléatoires indépendantes de lois respectives $\chi^2(n_1)$ et $\chi^2(n_2)$. Alors par définition:

$$F = \frac{S_1/n_1}{S_2/n_2}$$

suit une loi de Fisher à n_1 et n_2 degrés de liberté, notée $F(n_1, n_2)$.

Remarque. Par les mêmes considérations que précédemment, la loi F est d'autant plus proche de 1 que les degrés de liberté n_1 et n_2 sont grands.

On a également les propriétés suivantes:

- Si F suit une loi $F(n_1, n_2)$, alors la loi de $\frac{n_1}{n_2} F$ est une loi beta de seconde espèce de paramètres $(n_1/2, n_2/2)$, c'est-à-dire que F est à valeurs dans $\mathbb{R}_+$ et la densité par rapport à la mesure de Lebesgue:

$$f_X(x) = \frac{1}{B(n_1/2, n_2/2)} n_1^{n_1/2} \cdot n_2^{n_2/2} \frac{x^{n_1/2-1}}{(n_2 + n_1 \cdot x)^{(n_1+n_2)/2}} \mathbb{1}_{\{x \geq 0\}},$$

  la notation Beta désignant encore la fonction Beta.
- Si $F \sim F(n_1, n_2)$, alors $\mathbb{E}(F) = \frac{n_2}{n_2 - 2}$ lorsque $n_2 > 2$ et $\mathrm{var}(F) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2}$ lorsque $n_2 > 4$.
- Si T suit une loi de Student T (n), alors T^2 suit une loi de Fisher F (1 , n).

La figure A.4 donne une idée de la distribution d'une loi de Fisher pour différents choix des paramètres.

2.4 Indépendance

Définitions. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé.

– Soit $(A_i)_{i \in I}$ une famille dénombrable d'événements de $\mathcal{A}$. On dit que les événements $(A_i)_{i \in I}$ sont indépendants si et seulement si pour toute sous-ensemble fini K de I,

$$\mathbb{P}\left(\bigcap_{i \in K} A_i\right) = \prod_{i \in K} \mathbb{P}(A_i).$$

– Soit $(\mathcal{A}_i)_{i \in I}$ une famille de sous-tribus de $\mathcal{A}$ (donc pour tout $i \in I$, $\mathcal{A}_i \subset \mathcal{A}$). On dit que les tribus $(\mathcal{A}_i)_{i \in I}$ sont indépendantes si et seulement si pour tout sous-ensemble fini K de I, pour tous les événements $A_k \in \mathcal{A}_k$ avec $k \in K$, les A_k sont indépendants.

– Soit $(X_i)_{i \in I}$ des variables aléatoires sur $(\Omega, \mathcal{A})$, à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. On dit que les v.a. $(X_i)_{i \in I}$ sont indépendantes si et seulement si les tribus engendrées $(X_i^{-1}(\mathcal{B}(\mathbb{R})))_{i \in I}$ sont indépendantes.

Proposition. Si $(X_1, \cdots, X_n)$ sont des variables aléatoires sur $(\Omega, \mathbb{P})$. Alors les (X_i) sont indépendantes si et seulement si $\mathbb{P}_{(X_1, \cdots, X_n)} = \bigotimes_{i=1}^{n} \mathbb{P}_{X_i}$.

Proposition. Si $(X_i)_{i \in I}$ sont des variables aléatoires indépendantes sur $(\Omega, \mathbb{P})$. Alors les (X_i) sont indépendantes si et seulement si pour toute famille finie, pour toutes fonctions mesurables g_j, $j \in J$ telles que $g_j(X_j)$ soit intégrable, alors

$$\mathbb{E}\left[\prod_{j \in J} g_j(X_j)\right] = \prod_{j \in J} \mathbb{E}\left(g_j(X_j)\right).$$

Corollaire. $(X_1, \cdots, X_n)$ sont des variables aléatoires indépendantes si et seulement si pour tout $t = (t_1, \cdots, t_n) \in \mathbb{R}^n$,

$$\varphi_{(X_1, \cdots, X_n)}(t_1, \cdots, t_n) = \prod_{j=1}^{n} \varphi_{X_j}(t_j).$$

2.5 Vecteurs aléatoires

Définition. On dit que X est un vecteur aléatoire sur $(\Omega, \mathbb{P})$, un espace probabilisé, si X est une fonction mesurable de Ω dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

Définition. Soit X un vecteur aléatoire sur $(\Omega, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$. Alors la loi (ou mesure) de probabilité de X, $\mathbb{P}_X$, est définie de façon unique à partir de la fonction de répartition de X, telle que pour $x = (x_1, \cdots, x_d)$,

$$F_X(x) = \mathbb{P}_X\left(\prod_{i=1}^{d}]-\infty, x_i]\right) = \mathbb{P}\left(X \in \prod_{i=1}^{d}]-\infty, x_i]\right).$$

Propriété. Soit X un vecteur aléatoire sur $(\Omega, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$ et supposons que $X = (X_1, \cdots, X_d)$. Alors les X_i sont des variables aléatoires sur $(\Omega, \mathbb{P})$, de fonction de répartition

$$F_{X_i}(x_i) = \lim_{\substack{x_j \to +\infty \\ j \neq i}} F_X(x_1, \cdots, x_i, \cdots, x_d).$$

Les mesures de probabilité définies de façon unique à partir de ces F_{X_i} sont appelées lois marginales de X.

On se place maintenant dans la base canonique orthonormée de $\mathbb{R}^d$. Si Z est un vecteur aléatoire à valeurs sur $\mathbb{R}^d$, on définit $\mathbb{E}(Z)$, le vecteur dont les coordonnées sont les espérances des coordonnées de Z. Ainsi, si dans la base canonique de $\mathbb{R}^d$, $Z = (Z_1, \cdots, Z_d)'$,

$$\mathbb{E}(Z) = \mathbb{E}\begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}(Z_1) \\ \vdots \\ \mathbb{E}(Z_d) \end{pmatrix}.$$

De la même manière, on définit l'espérance d'une matrice dont les coefficients sont des variables aléatoires par la matrice dont les coefficients sont les espérances de chacune de ces variables aléatoires.

Ceci nous permet de définir la matrice de variance-covariance de Z de la manière:

$$\mathrm{var}(Z) = \mathbb{E}\left[(Z - \mathbb{E}(Z))(Z - \mathbb{E}(Z))'\right]$$

donc si $Z = (Z_1, \cdots, Z_d)'$,

$$\mathrm{var}\begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} = \begin{pmatrix} \mathrm{var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdots & \mathrm{Cov}(Z_1, Z_d) \\ \mathrm{Cov}(Z_1, Z_2) & \mathrm{var}(Z_2) & \cdots & \mathrm{Cov}(Z_2, Z_d) \\ \vdots & \vdots & \cdots & \vdots \\ \mathrm{Cov}(Z_1, Z_d) & \mathrm{Cov}(Z_2, Z_d) & \cdots & \mathrm{var}(Z_d) \end{pmatrix}$$

matrice (d, d) dont les éléments diagonaux sont les variances des Z_i et les éléments diagonaux sont les covariances des coordonnées de Z (remarquons que la variance est celle de Z_i si la covariance est celle de Z_i et Z_1).

On vérifie également aisément que si C est une matrice (p, d) à coefficients constants et si Z est un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, alors CZ est un vecteur de taille p de matrice de variance-covariance

$$\mathrm{var}(CZ) = C \cdot \mathrm{var}(Z) \cdot C'.$$

En particulier, si p vaut 1, alors $C = h'$ est un vecteur de taille d, et:

$$\mathrm{var}(h'Z) = h' \cdot \mathrm{var}(Z) \cdot h.$$

Notez que cette dernière quantité est un scalaire. Soit $Y_1, \cdots, Y_d$ des variables aléatoires indépendantes de même loi $N(0, \sigma^2)$, indépendantes (ce qui, dans le cas gaussien, est équivalent à $\mathrm{cov}(Y_i, Y_j) = 0$ pour $i = j$). On considère le vecteur $Y = (Y_1, \cdots, Y_d)'$. En raison de l'indépendance, Y est un vecteur gaussien admettant

une densité f_Y (par rapport à la mesure de Lebesgue d) sur $\mathbb{R}^d$ est le produit des densités de chacune des coordonnées, soit:

$$\begin{aligned} f_Y(y_1,\cdots,y_d) &= f_{Y_1}(y_1)\times f_{Y_2}(y_2)\times\cdots\times f_{Y_d}(y_d) \\ &= (2\pi\sigma^2)^{-d/2}\exp\left(-\frac{1}{2\sigma^2}(y_1^2+\cdots+y_d^2)\right) \\ &= (2\pi\sigma^2)^{-d/2}\exp\left(-\frac{\|y\|^2}{2\sigma^2}\right), \end{aligned}$$

avec $y=(y_1,\cdots,y_d)$. On voit donc que la densité ne dépend que de la norme $\|y\|$: elle est constante sur toutes les sphères centrées en 0. Cela implique qu'elle est invariante par rotation autour de l'origine: elle est invariante par $O(d)$ tout entier, et pas seulement par $SO(d)$. On dira que Y suit une loi gaussienne isotrope. Rappelons que les densités correspondent à des changements de base orthonormée. En conséquence, on a la première propriété importante:

Propriété. Soit Y un vecteur aléatoire de $\mathbb{R}^d$ de loi normale isotrope variance σ^2, c'est-à-dire que dans une BON les coordonnées de Y vérifient $\mathbb{E}(Y)=0$ et $\mathrm{var}(Y)=\sigma^2\cdot Id$. Alors les coordonnées de Y dans toute BON sont encore des $N(0,\sigma^2)$ indépendantes.

Voici maintenant l'un des résultats (encore appelé Théorème de Cochran) que nous utiliserons le plus et nous en donnons donc une démonstration.

Théorème (Théorème de Cochran). Soit E_1 et E_2, deux sous-espaces vectoriels orthogonaux de $E=\mathbb{R}^d$ de dimensions respectives k_1 et k_2 et soit Y un vecteur aléatoire de $\mathbb{R}^d$ de loi normale centrée et isotrope de variance σ^2. Alors $P_{E_1}(Y)$ et $P_{E_2}(Y)$ sont deux variables aléatoires gaussiennes centrées indépendantes et $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2\chi^2(k_1)$ (resp. $\sigma^2\cdot\chi^2(k_2)$). Ce théorème se généralise naturellement pour 2 remplacé par $m\geq 2$ sous-espaces vectoriels orthogonaux de $E=\mathbb{R}^d$.

Démonstration: Soit $(e_1,\cdots,e_{k_1})$ et $(e_{k_1+1},\cdots,e_{k_1+k_2})$ des BON de E_1 et E_2 (respectivement). L'ensemble de ces deux bases peut être complété en

$$(e_1,\cdots,e_{k_1},e_{k_1+1},\cdots,e_{k_1+k_2},e_{k_1+k_2+1},\cdots,e_d)$$

pour former une BON de $\mathbb{R}^d$ (du fait que E_1 et E_2 sont orthogonaux).

Soit $(Y_1,\cdots,Y_d)$, les coordonnées de Y dans cette base. Elles sont indépendantes de loi $N(0,\sigma^2)$ car le changement de base est orthonormal et nous avons vu que la distribution gaussienne centrée Y transformation isométrique. Comme

$$\begin{aligned} P_{E_1}(Y)=Y_1e_1+\cdots+Y_{k_1}e_{k_1} \quad&\Longrightarrow\quad \|P_{E_1}(Y)\|^2=\sigma^2\left(\frac{Y_1}{\sigma}\right)^2+\cdots+\left(\frac{Y_{k_1}}{\sigma}\right)^2 \\ P_{E_2}(Y)=Y_{k_1+1}e_{k_1+1}+\cdots+Y_{k_1+k_2}e_{k_1+k_2} \quad&\Longrightarrow\quad \|P_{E_2}(Y)\|^2=\sigma^2\left(\frac{Y_{k_1+1}}{\sigma}\right)^2+\cdots+\left(\frac{Y_{k_1+k_2}}{\sigma}\right)^2. \end{aligned}$$

On voit bien ainsi l'indépendance entre ces deux projections et le fait que la loi de $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2\cdot\chi^2(k_1)$ (resp. $\sigma^2\cdot\chi^2(k_2)$). ■

On peut définir plus généralement un vecteur gaussien Y à valeurs dans $\mathbb{R}^d$ (ou dans $\mathbb{R}$ généré), de espérance $\mu\in\mathbb{R}^d$ et de matrice de variance-covariance Σ quelconques (du moment que Σ soit une matrice de Toeplitz définie positive). Cela va utile à définir un vecteur gaussien de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$,

$$f_Y(y)=\frac{(2\pi)^{-n/2}}{|\Sigma|}\exp\left(-\frac{1}{2}(y-\mu)'.\Sigma^{-1}.(y-\mu)\right),$$

pour $y\in\mathbb{R}^d$, et avec $|\Sigma|$ le déterminant de la matrice Σ. Remarquons une nouvelle fois que les paramètres μ et Σ contiennent la loi de probabilité du vecteur gaussien.

A partir des propriétés énoncées sur les vecteurs gaussiens, on obtient le fait que:

Propriété. Soit Y un vecteur gaussien à valeurs dans IR^d (non dégénéré), d'espérance $\mu \in IR^d$ et de matrice de variance-covariance Σ. Soit C une matrice réelle de taille (p, d) où $p \in IN^*$. Alors $C \cdot Y$ est un vecteur gaussien tel que:

$$C \cdot Y \sim N(C \cdot \mu, C \cdot \Sigma \cdot C')$$

On en déduit les conséquences suivantes:

- si Y est un vecteur gaussien isotrope de IR^d de variance σ^2 et h un vecteur de IR^d, alors $h' \cdot Y$ est une combinaison linéaire des coordonnées de Y telle que:

$$h' \cdot Y \text{ suit la loi } N(0, \sigma^2 \cdot h' \cdot h) = N(0, \sigma^2 \cdot \|h\|^2)$$

- si Y est un vecteur gaussien d'espérance μ et de matrice de variance Σ et si h un vecteur de IR^d, alors $h' \cdot Y$ est une combinaison linéaire des coordonnées de Y et:

$$h' \cdot Y \text{ suit la loi unidimensionnelle } N(h' \cdot \mu, h' \cdot \Sigma \cdot h)$$

(Pour une présentation plus détaillée des notions sur les vecteurs gaussiens on peut consulter le livre P. Toulouse, 1999, chap.2)

2.6 Fonctions caractéristiques

Définition. Soit X un vecteur aléatoire sur (Ω, A, IP) à valeurs dans IR^d. La fonction caractéristique de X est la fonction $\varphi_X : IR^d \to \mathbb{C}$ telle que

$$\varphi_X(t) = IE[\exp(i < t, X >)] = \int_{IR^d} e^{i<t,x>} dIP_X(x),$$

où <. > désigne le produit scalaire euclidien sur IR^d tel que $< t, x > = \sum_{i=1}^{d} t_i x_i$ pour $t = (t_1, \cdots, t_d)$ et $x = (x_1, \cdots, x_d)$.

Remarque. La fonction caractéristique existe sur IR^d et $\varphi_X(0) = 1$. φ_X est aussi la transformée de Fourier de la mesure IP_X.

Théorème. Soit X et Y des vecteurs aléatoires sur (Ω, A, IP) à valeurs dans IR^d, de lois IP_X et IP_Y. Alors $IP_X = IP_Y$ si et seulement si $\varphi_X = \varphi_Y$.

Théorème (Théorème d'inversion). Si X est un vecteur aléatoire sur (Ω, A, IP) à valeurs dans IR^d et si φ_X est une fonction intégrable par rapport à la mesure de Lebesgue sur IR^d, alors X admet une densité f_X par rapport à λ_d telle que pour $x \in IR^d$,

$$f_X(x) = \frac{1}{(2\pi)^d} \int_{IR^d} e^{-i<t,x>} \varphi_X(t) dt.$$

Proposition. Si X est une variable aléatoire sur (Ω, A, IP) de fonction caractéristique φ_X. Alors si $IE(|X|^n) < +\infty$ (ou $X \in IL^n(\Omega, A, IP)$), φ_X est n fois dérivable et $\varphi_X^{(n)}(t) = i^n IE(X^n e^{itX})$.

Remarque. Lorsque ces moments existent, on a $i^n IE(X^n) = \varphi_X^{(n)}(0)$.

2.7 Convergence de suites de variables aléatoires

Lemme (Lemme de Borel-Cantelli). Soit $(A_n)_{n \in IN}$ une suite d'événements sur (Ω, A, IP).

1. Si $\sum IP(A_n) < +\infty$ alors $IP(\limsup A_n) = 0$.
2. Si les (A_n) sont indépendants, $\sum IP(A_n) = +\infty$ implique que $P(\limsup A_n) = 1$.

Définition. Soit $(X_n)_{n \in IN}$ une suite de variables aléatoires sur (Ω, A, IP). On dit que

– (X_n) converge en probabilité vers X, noté $X_n \xrightarrow[n\to+\infty]{P} X$, lorsque pour tout ε > 0,

$$\lim_{n\to\infty} IP(|X_n - X| > \varepsilon) = 0.$$

– (X_n) converge dans $L^p(\Omega, \mathcal{A}, \mathbb{P})$ vers X, noté $X_n \xrightarrow[n\to+\infty]{L^p} X$, avec $p > 0$, lorsque

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0.$$

– (X_n) converge en loi vers X, noté $X_n \xrightarrow[n\to+\infty]{L} X$, lorsque,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \text{pour tout } x \in \mathbb{R} \text{ tel que } F_X \text{ continue en } x.$$

– (X_n) converge presque sûrement vers X, noté $X_n \xrightarrow[n\to+\infty]{p.s.} X$, lorsque $\exists E \in \mathcal{A}$ avec $P(E) = 1$ tel que
$\forall \omega \in E$, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$
$\Leftrightarrow$ pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{m \geq n} |X_m - X| > \varepsilon\right) = 0.$$

Propriété.
1. p.s. et $L^p \Longrightarrow P \Longrightarrow L$.
2. pour $q \geq p$, $L^q \Longrightarrow L^p$.
3. La convergence en loi n'implique pas la convergence en probabilité. Mais si $(X_n \xrightarrow[n\to+\infty]{P} C) \Longleftrightarrow (X_n \xrightarrow[n\to+\infty]{L} C)$ pour C une constante.
4. Si g est une fonction borélienne continue alors $(X_n \xrightarrow[n\to+\infty]{P} X) \Rightarrow (g(X_n) \xrightarrow[n\to+\infty]{P} g(X))$.

Propriété.
1. Si pour tout $\varepsilon > 0$, $\sum_{n=0}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < +\infty$ alors $X_n \xrightarrow[n\to+\infty]{p.s.} X$ (application du Lemme de Borel-Cantelli).
2. Si il existe $r > 0$ tel que $\mathbb{E}(|X_n|^r) < +\infty$ et $\sum_{n=0}^{\infty} \mathbb{E}(|X_n - X|^r) < +\infty$ alors $X_n \xrightarrow[n\to+\infty]{p.s.} X$.

Théorème (Loi faible des Grands Nombres). Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées. Alors si $\mathbb{E}(|X_i|) < +\infty$,

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to+\infty]{P} m = \mathbb{E}X_i.$$

Théorème (Loi forte des Grands Nombres). Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées. Alors si $\mathbb{E}(|X_i|) < +\infty$,

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to+\infty]{p.s.} m = \mathbb{E}X_i.$$

Théorème (Théorème de la limite centrale). Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées. Alors si $\sigma^2 = \mathbb{E}X_i^2 < +\infty$, et $m = \mathbb{E}X_i$,

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \xrightarrow[n\to+\infty]{L} \mathcal{N}(0, 1).$$

Théorème (Loi forte des Grands Nombres multidimensionnelle). Soit $(X_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$, indépendants et identiquement distribués. Alors si $\mathbb{E}(X_i) < +\infty$ (pour $\|\cdot\|$ une norme sur $\mathbb{R}^d$),

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to+\infty]{p.s.} m = \mathbb{E}X_i.$$

Théorème (Théorème de la limite centrale multidimensionnelle). Soit $(X_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$, indépendants et identiquement distribués. Alors si Σ matrice de covariance de chaque X_i existe, et $m = \mathbb{E}X_i$,

$$\sqrt{n}\overline{X}_n - m \xrightarrow[n\to+\infty]{L} \mathcal{N}_d(0, \Sigma).$$

Théorème (Delta-method). Soit $(X_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$, indépendants et identiquement distribués, tel que Σ matrice de covariance de chaque X_i existe, et $m = \mathbb{E}X_i$. Soit $g : \mathbb{R}^d \to \mathbb{R}^p$ une fonction de classe C^1 sur un voisinage autour de m, de matrice Jacobienne J_g en m. Alors,

$$\sqrt{n}\,g(\overline{X}_n) - g(m) \xrightarrow[n\to+\infty]{L} \mathcal{N}_d(0, J_g(m) \cdot \Sigma \cdot J_g(m)).$$

2.8 Espérance conditionnelle

Définitions. Soit Y une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$. Si $\mathcal{B}$ est une sous-tribu de $\mathcal{A}$ et si $Y \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P})$. Alors on note $\mathbb{E}(Y|\mathcal{B})$ la projection orthogonale de Y sur $\mathbb{L}^2(\Omega, \mathcal{B}, \mathbb{P})$, appelée espérance conditionnelle de Y sachant $\mathcal{B}$. Ainsi:

$$\mathbb{E}\,|Y - \mathbb{E}(Y|\mathcal{B})|^2 = \inf_{Z \in \mathbb{L}^2(\Omega, \mathcal{B}, \mathbb{P})} \mathbb{E}\,|Y - Z|^2 \quad .$$

Par extension, si $Y \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, on définit l'espérance conditionnelle par rapport à $\mathcal{B}$ comme l'unique (p.s.) variable aléatoire $\mathcal{B}$-mesurable vérifiant p.s.:

$$\int_B \mathbb{E}(Y|\mathcal{B})\, d\mathbb{P} = \int_B Y\, d\mathbb{P}, \quad \text{pour tout } B \in \mathcal{B}.$$

Définition. Par convention, si X un vecteur aléatoire à valeurs dans $\mathbb{R}^n$ sur $(\Omega, \mathcal{A}, \mathbb{P})$ et si Y une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, on note $\mathbb{E}(Y|X) = \mathbb{E}(Y|X^{-1}(\mathcal{B}(\mathbb{R})))$.

Propriété.

1. Lemme de Doob : Pour $Y \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, et X une v.a. de $(\Omega, \mathcal{A}, \mathbb{P})$, alors p.s. $\mathbb{E}(Y|X) = h(X)$, avec h une fonction borélienne.
2. Pour Y_1 et Y_2 deux variables aléatoires sur $(\Omega, \mathcal{A}, \mathbb{P})$, et $(a, b, c) \in \mathbb{R}^3$, alors
$$\mathbb{E}(aY_1 + bY_2 + c\,|\mathcal{B}) = a\mathbb{E}(Y_1|\mathcal{B}) + b\mathbb{E}(Y_2|\mathcal{B}) + c.$$
3. Si $Y_1 \le Y_2$, alors $\mathbb{E}(Y_1|\mathcal{B}) \le \mathbb{E}(Y_2|\mathcal{B})$.
4. Le Lemme de Fatou, les théorèmes de Beppo-Levi, Lebesgue et Jensen s'appliquent avec l'espérance conditionnelle.
5. Si $Y \in \mathbb{L}^2(\Omega, \mathcal{B}, \mathbb{P})$, alors $\mathbb{E}(Y|\mathcal{B}) = Y$; ainsi $\mathbb{E}(g(X)\,|X) = g(X)$ pour g une fonction mesurable réelle.
6. On a $\mathbb{E}(\mathbb{E}(Y|\mathcal{B})) = \mathbb{E}\,Y$.
7. Si $Y^{-1}(\mathcal{B}(\mathbb{R}))$ et $\mathcal{B}$ sont indépendantes alors $\mathbb{E}(Y|\mathcal{B}) = \mathbb{E}Y$; ainsi si X et Y sont indépendantes, $\mathbb{E}(Y|X) = \mathbb{E}Y$.
8. Si (X, Y) est un couple de v.a. à valeurs dans $\mathbb{R}^2$ possédant une densité $f_{(X,Y)}$ par rapport à la mesure de Lebesgue, alors si X est intégrable,
$$\mathbb{E}(Y \mid X = x) = \frac{\int_{\mathbb{R}} y \cdot f_{(X,Y)}(x, y)\, dy}{\int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dy}, \quad \text{pour tout } x \text{ tel que } \int_{\mathbb{R}} f_{(X,Y)}(x, y)dy > 0.$$

Proposition. Si $(Y, X_1, \cdots, X_n)$ est un vecteur gaussien, alors $\mathbb{E}(Y|X_1, \cdots, X_n)) = a_0 + a_1X_1 + \cdots + a_nX_n$, où les a_i sont des réels.

3 Estimation paramétrique

3.1 Définitions

Dans toute la suite, on se place sur $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. On dit que $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires, où chaque X_n est définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ et est à valeur dans $\mathbb{R}$.

Définition.
- On appelle modèle statistique de dimension n un espace $(\Omega', \mathcal{A}'_n, \mu)$, où $\mathcal{A}'_n$ est une tribu sur Ω'^n et μ une mesure de probabilité sur $(\Omega'^n, \mathcal{A}'_n)$.
- On appelle échantillon de taille n du modèle statistique $(\Omega'^n, \mathcal{A}'^n_n, \mu)$ le vecteur aléatoire $(X_1, \ldots, X_n)$ distribué selon la loi μ. Pour $\omega \in \Omega$, $(X_1(\omega), \ldots, X_n(\omega))$ vecteur de $\mathbb{R}^n$ est appelé échantillon observé. C'est à partir de ce vecteur qu'est statistique s'effectue généralement.

Définition. On appelle:
- Modèle statistique paramétrique, une famille de modèle de la forme : $(\Omega'^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$.
- Modèle statistique semi-paramétrique, une famille de modèle de la forme : $(\Omega'^n, \mathcal{A}'_n, \mathbb{P}_{(\theta,f)}, \theta \in \Theta, f \in F)$, où $\Theta \subset \mathbb{R}^p$ et F n'est pas de dimension finie.

– Modèle statistique non-paramétrique, une famille de modèle de la forme : $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_f, f \in \mathcal{F})$, où $\mathcal{F}$ n'est pas de dimension finie.

Définition. – On dit que le modèle paramétrique : $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, est dominé par une mesure μ lorsque $\mathbb{P}_\theta$ est absolument continue par rapport à μ pour tout $\theta \in \Theta$.

– On se place dans le cadre d'un modèle paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure μ. Pour $(x_1, \cdots, x_n) \in (\Omega')^n$, la fonction $\theta \in \Theta \to L_\theta(x_1, \cdots, x_n) = \frac{d\mathbb{P}_\theta}{d\mu}(x_1, \cdots, x_n)$ est appelée une vraisemblance du modèle statistique.

Exemple. – Dans le cas où μ est la mesure de Lebesgue sur $\mathbb{R}^n$, la vraisemblance sera la densité (classique) en $(x_1, \cdots, x_n)$.

– Dans le cas où μ est comptage sur $\mathbb{N}^n$, la vraisemblance sera la probabilité en $(x_1, \cdots, x_n)$.

– Attention ! si le support de $\mathbb{P}_\theta$ dépend de θ, la mesure qui domine (ainsi que $\mathcal{A}'_n$ et Ω') peut dépendre de θ : il ne faut pas oublier de le préciser dans l'expression de la vraisemblance.

Définition. Lorsque l'on dispose d'un échantillon $(X_1, \ldots, X_n)$ du modèle statistique $((\Omega')^n, \mathcal{A}'_n, \mu)$, une statistique T_n est une application mesurable de $((\Omega')^n, \mathcal{A}'_n)$ dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, donc un vecteur aléatoire défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeur dans $\mathbb{R}^d$ et tel que:

$$T_n = h(X_1, \ldots, X_n), \quad \text{où } h : (\Omega')^n \to \mathbb{R}^d \text{ est mesurable.}$$

Exemple. Estimateur du paramètre d'une loi de Bernoulli.
Estimateur de l'espérance et de la variance par la moyenne et la variance empirique.
Estimateurs du paramètre θ d'un échantillon $(X_1, \cdots, X_n)$ de loi uniforme sur $[0, \theta]$.
Test sur la moyenne.

3.2 Statistiques exhaustives

On se place désormais dans le cadre d'une modélisation statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominée par une mesure μ.

Exemple. 1. Soit le modèle statistique paramétrique $([0, \infty[^n, \mathcal{B}([0, \infty[^n), \mathcal{U}([0, \theta])^{\otimes n}, \theta \in]0, +\infty[)$. On dispose donc d'un échantillon $(X_1, \cdots, X_n)$ de v.a.i.i.d. suivant une loi uniforme sur $[0, \theta]$. Si on considère $\max\{X_1, \ldots, X_n\}$ cela semble suffire pour récupérer toute l'information sur θ que contenait $(X_1, \ldots, X_n)$: on a donc résumé l'"information" sur θ contenait $(X_1, \ldots, X_n)$, un vecteur de taille n, par une statistique de taille 1.
2. De même, si on considère le modèle statistique paramétrique $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathcal{B}(p)^{\otimes n}, p \in [0, 1])$ (on dispose donc d'un échantillon $(X_1, \cdots, X_n)$ de v.a.i.i.d. suivant une loi de Bernoulli de paramètre p) alors la statistique $X_1 + \cdots + X_n$ contient toute l'"information" sur p contenue dans l'échantillon $(X_1, \ldots, X_n)$.

Comment exprimer formellement le fait qu'une statistique puisse renfermer à elle seule toute l'information sur le paramètre?

Définition. Soit T une statistique d'un modèle statistique paramétrique dominé à valeurs dans $\mathbb{R}^d$. On dit que T est une statistique exhaustive si pour toute statistique S intégrable (donc dans $\mathbb{L}^1((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta)$) alors $\mathbb{E}_\theta(S \mid T)$ ne dépend ($\mathbb{P}_\theta$ presque sûrement) pas de θ.

Théorème (Théorème de factorisation de Neyman). Soit $(X_1, \cdots, X_n)$ un n-échantillon et soit T une statistique d'un modèle statistique paramétrique dominé avec T à valeurs dans $\mathbb{R}^d$, où $d \in \mathbb{N}^*$. La statistique T est exhaustive si et seulement s'il existe une fonction $h(.) : \mathbb{R}^n \to \mathbb{R}_+$ et une fonction $g_\theta(.) : \mathbb{R}^d \to \mathbb{R}_+$, telle que l'on puisse écrire pour tout $(x_1, \ldots, x_n) \in (\Omega')^n$:

$$L_\theta(x_1, \ldots, x_n) = g_\theta(T(x_1, \ldots, x_n)) \cdot h(x_1, \ldots, x_n) \quad \text{pour tout } \theta \in \Theta.$$

Lemme. Soit le modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$. Alors ce modèle est dominé si et seulement si il existe une sous-famille dénombrable $(\mathbb{P}_{\theta_i})_{i \in \mathbb{N}}$ telle que pour tout $A \in \mathcal{A}$, $\forall i \in \mathbb{N}$, $\mathbb{P}_{\theta_i}(A) = 0$ entraîne $\forall \theta$, $\mathbb{P}_\theta(A) = 0$. Toute mesure de probabilité de la forme $\mathbb{P}^* = \sum_{i \in \mathbb{N}} c_i \cdot \mathbb{P}_{\theta_i}$ avec $c_i > 0$ pour tout $i \in \mathbb{N}$ et $\sum_{i \in \mathbb{N}} c_i = 1$ domine le modèle.

Démonstration du lemme : $\Longrightarrow$ Il est bien clair que si une telle mesure P^* existe, le modèle est dominé.
$\Longrightarrow$ Montrons maintenant que si le modèle est dominé par une mesure μ alors la famille $(P_{\theta_i})_{i\in\mathbb{N}}$ existe.
En premier lieu, si μ est une mesure non finie mais σ-finie (par exemple la mesure de Lebesgue), définie par $\mu(A)=\sum_{i=1}^{\infty}\frac{1}{2^i}\frac{\mu(A\cap A_i)}{\mu(A_i)}$ pour tout $A\in\mathcal{A}$, est une mesure de probabilité équivalente à μ (avec $(A_i)_{i\in\mathbb{N}^*}$ une partition de Ω telle que $0<\mu(A_i)<\infty$ pour tout $i\in\mathbb{N}^*$). On travaille donc désormais avec μ.
Pour $\theta\in\Theta$, soit B_θ le sous-ensemble de $(\Omega^n)^n$ qui est le support de la densité P_θ par rapport à μ. Soit

$$C=\bigcup_{i\in I} B_{\theta_i},\ I\subset\mathbb{N},\theta_i\in\Theta,$$

l'ensemble de toutes les unions dénombrables d'ensembles B_θ. Notons $M=\sup_{C\in\mathcal{C}}\mu_P(C)$. Soit $(C_n)_{n\in\mathbb{N}}$ une suite d'ensembles de $\mathcal{C}$ telle que la suite $(\mu_P(C_n))_n$ converge vers M (une telle suite existe forcément sinon M ne serait pas le supremum). Remarquons que, pour chaque C_n, énumérable, on a alors une suite (θ_k) de Θ suffit pour engendrer la suite $(C_n)_{n\in\mathbb{N}}$. Si on pose :

$$D=\bigcup_{n\in\mathbb{N}} C_n=\bigcup_{k\in\mathbb{N}} B_{\theta_k},$$

alors $M=\mu_P(D)$ et pour tout $\theta\in\Theta, B_\theta\cup D\in\mathcal{C}$ et :

$$\mu_P(B_\theta\cup D)\le M=\mu_P(B_\theta\cup D)=\mu_P(B_\theta\cap D^c)+\mu_P(D)$$

Donc pour tout $\theta\in\Theta,\mu_P(B_\theta\cap D^c)=0$ soit $\forall\theta\in\Theta, P_\theta(B_\theta\cap D^c)=0$ puisque $P_\theta\ll\mu_P$. En conséquence, pour tout $A\in\mathcal{A}', A\subset B_\theta\cup B_\theta^c=(\Omega')^n$, soit :

$$P_\theta(A\cap D^c)=0,\quad \text{car par définition des } B_\theta,\ P_\theta(B_\theta^c)=0.$$

Si on suppose maintenant que $A\in\mathcal{A}'$ est tel que $P_{\theta_k}(A)=0$, avec (P_{θ_k}) précédemment définie, alors $\mu_P(A\cap B_{\theta_k})=0$ par définition des B_θ et donc $\mu_P(A\cap D)=0$ (par la propriété σ-additivité d'une mesure).
Comme $P_\theta\ll\mu_P$, on en déduit que $\forall\theta\in\Theta, P_\theta(A\cap D)=0$ et donc $P_\theta(A)=P_\theta(A\cap D)+=P_\theta(A\cap D^c)=0$.
Ainsi, P^* domine bien P_θ pour tout $\theta\in\Theta$. ■

Démonstration du théorème de factorisation de Neyman : Soit $P^*=\sum_{i\in\mathbb{N}}a_i\cdot P_{\theta_i}$ une mesure de probabilité dominante construite comme dans le lemme.
$\Longleftarrow$ Si $g_\theta(T(x))\cdot h(x)$ avec $x\in(\Omega')^n$ est la densité P_θ par rapport à μ, alors $\sum_{i\in\mathbb{N}}a_i\cdot g_{\theta_i}(T(x))\cdot h(x)=g_*(T(x))\cdot h(x)$ est une densité P^* par rapport à μ. Alors, comme $g_*(T(x))\cdot h(x)>0 P^*$-p.s., donc P_θ-p.s.,
pour toute variable aléatoire S intégrable, pour tout $\theta\in\Theta$

$$\begin{aligned}
E_\theta(S\cdot 1\!\!1_B) &= \int_B S\,dP_\theta,\quad \text{pour tout } B\in\sigma(T),\ \text{tribu engendrée par } T\\
&= \int_B S(x)\cdot g_\theta(T(x))\cdot h(x)\,d\mu(x)\\
&= \int_B S(x)\cdot\frac{g_\theta(T(x))\cdot h(x)}{g_*(T(x))\cdot h(x)}\,dP^*(x)\\
&= E_*\left(1\!\!1_B\cdot\frac{g_\theta(T)}{g_*(T)}\cdot S\right)\\
&= E_*\left(1\!\!1_B\cdot\frac{g_\theta(T)}{g_*(T)}\cdot E_*(S\mid T)\right)\quad \text{(d'après la définition de l'espérance conditionnelle)}\\
&= E_\theta\left(1\!\!1_B\cdot E_*(S\mid T)\right).
\end{aligned}$$

En conséquence, d'après la définition de l'espérance conditionnelle dans $(\Omega',\mathcal{A}',P_\theta)$, on a P_θ- p.s.,
$E_*(S\mid T)=E_\theta(S\mid T)$: la statistique T est bien exhaustive.
$\Longrightarrow$ On suppose que T est une statistique exhaustive pour le modèle. Donc pour toute statistique intégrable

S, $\forall\theta$, $\mathbb{E}_\theta(S \mid T) = \mathbb{E}_\theta(S \mid T)$. En conséquence, si on note $\varphi(x,\theta) = \frac{d\mathbb{P}_\theta}{d\mathbb{P}_*}(x)$ la densité de $\mathbb{P}_\theta$ par rapport à $\mathbb{P}_*$,

$$\begin{aligned}
\mathbb{E}_\theta(S) &= \mathbb{E}_\theta\, \mathbb{E}_*(S \mid T), \quad \text{(car } T \text{ est exhaustive et d'après le théorème de l'espérance conditionnelle)}\\
&= \mathbb{E}_* \varphi(X,\theta)\, \mathbb{E}_\mu(S \mid T), \quad \text{où } X \sim \mathbb{P}_*\\
&= \mathbb{E}_* \mathbb{E}_*\big[\varphi(X,\theta)\, \mathbb{E}_*(S \mid T) \mid T\big], \quad \text{(d'après le théorème de l'espérance conditionnelle)}\\
&= \mathbb{E}_* \mathbb{E}_*\big[\varphi(X,\theta) \mid T\big] \cdot \mathbb{E}_*\big[S \mid T\big], \quad \text{(car } \mathbb{E}_*[S \mid T] \text{ est une fonction de } T)\\
&= \mathbb{E}_* \mathbb{E}_*\big[S \cdot \mathbb{E}_*[\varphi(X,\theta) \mid T] \mid T\big]\\
&= \mathbb{E}_* \big[S \cdot \mathbb{E}_*[\varphi(X,\theta) \mid T]\big]
\end{aligned}$$

Ainsi, la variable aléatoire $\mathbb{E}_*[\varphi(X,\theta) \mid T]$, qui est une fonction de T (et donc elle-même une fonction sur $(\Omega')^n$), est la densité $\mathbb{P}_\theta$ par rapport à $\mathbb{P}_*$. Par suite, la vraisemblance, qui est la densité $\mathbb{P}_\theta$ par rapport à μ, s'écrit:

$$L_\theta(x_1,\dots,x_n) = \frac{d\mathbb{P}_\theta}{d\mu}(x_1,\dots,x_n) = \frac{d\mathbb{P}_\theta}{d\mathbb{P}_*}(x_1,\dots,x_n) \cdot \frac{d\mathbb{P}_*}{d\mu}(x_1,\dots,x_n) = \mathbb{E}_*[\varphi(X,\theta) \mid T] \cdot h(x_1,\dots,x_n),$$

avec h une fonction mesurable. ■

Exemple. Différentes statistiques exhaustives pour les paramètres inconnus de loi uniforme, loi de Bernoulli, de loi gaussienne...

Propriété. On se place dans le cadre d'un modèle paramétrique dominé.

1. La statistique $T = (X_1,\dots,X_n)$ est exhaustive.
2. Si T est une statistique exhaustive, et s'il existe une fonction h telle qu'une autre statistique U vérifie $T = h(U)$, alors U est également exhaustive.

On vient de voir que l'on peut toujours trouver une statistique exhaustive (même par exemple). Comme on a un intérêt certain à avoir à voulir le "maximum d'information" dans une statistique lorsque le paramètre est dans $\mathbb{R}^d$, on aimerait savoir quelle dimension minimale peut avoir cette statistique. En particulier, si $d = 1$, peut-on toujours trouver une statistique exhaustive de taille 1 ? L'exemple suivant montre que ce n'est pas toujours le cas:

Exemple. Soit le modèle statistique $(\mathbb{R}^n, \mathcal{B}([0,\infty[^n), (\mathbb{P}_\theta)^{\otimes n}, \theta \in \mathbb{R}_+)$, où la densité $\mathbb{P}_\theta$ par rapport à la mesure de Lebesgue est $f_\theta(x) = \theta^2 e^{-\theta x} \mathbb{1}_{x \in [0,\infty[}$. Alors les statistiques $T_1 = \max(X_1,\dots,X_n)$ et $T_2 = X_1 + \dots + X_n$ ne sont pas chacune exhaustives alors que (T_1, T_2) est exhaustive. On pourra même montrer que cette statistique est de taille minimale...

Définition. Une statistique exhaustive T d'un modèle paramétrique dominé avec T est dite minimale si pour toute autre statistique exhaustive U est telle qu'il existe une fonction h vérifiant:

$$T = h(U).$$

Proposition. Soit un modèle paramétrique dominé et soit $L_\theta(x_1,\dots,x_n)$ sa vraisemblance. Alors T est une statistique exhaustive minimale pour θ lorsque $\forall (x_1,\dots,x_n) \in (\Omega')^n$ et $\forall (y_1,\dots,y_n) \in (\Omega')^n$,

$$\theta \to \frac{L_\theta(x_1,\dots,x_n)}{L_\theta(y_1,\dots,y_n)} \text{ ne dépend pas de } \theta \iff T(x_1,\dots,x_n) = T(y_1,\dots,y_n). \qquad (2)$$

Démonstration de la proposition. On suppose que (2) est vraie et on suppose (sans perte de généralité) que la variable est strictement positive. Soit $t \in \mathrm{Im}(T((\Omega')^n))$. Notons $x^{(t)} \in T^{-1}(\{t\}) \subset (\Omega')^n$. Alors $\forall x \in T^{-1}(\{t\})$, $T(x) = T(x^{(T(x))})$ et donc d'après (2),

$$h(x) = \frac{L_\theta(x)}{L_\theta(x^{(T(x))})} \text{ est indépendant de } \theta.$$

Posons $g_\theta(t) = L_\theta(x^{(t)})$. Alors $L_\theta(x) = g_\theta(T(x)) \cdot h(x)$. Comme ceci est vrai pour tout $x \in (\Omega')^n$, la statistique T est bien exhaustive.
Supposons maintenant que S est une autre statistique exhaustive. Alors par le théorème de factorisation de Neyman, il existe deux fonctions $g_\theta^{(s)}$ et $h^{(s)}$ (ne dépendant pas de θ) telles que pour tout $x \in (\Omega')^n$, $L_\theta(x) = g_\theta^{(s)}(S(x)) \cdot h^{(s)}(x)$. Ainsi pour tout $x \in (\Omega')^n$ et $y \in (\Omega')^n$ tels que $S(x) = S(y)$, alors:

$$\frac{L_\theta(x)}{L_\theta(y)} = \frac{g_\theta^{(s)}(S(x)) \cdot h^{(s)}(x)}{g_\theta^{(s)}(S(y)) \cdot h^{(s)}(y)} = \frac{h^{(s)}(x)}{h^{(s)}(y)}, \quad \text{qui est indépendant de } \theta.$$

Mais d'après (2) ceci n'est possible que si $T(x) = T(y)$. Donc T est une fonction de S et la statistique T est donc minimale. ■.

Qu'elle serait une sorte d'opposée de la notion de statistique exhaustive minimale ? Ce devrait être une statistique ne dépendant pas du paramètre, soit:

Définition. Une statistique T d'un modèle paramétrique est dite libre si sa loi ne dépend pas du paramètre.

Or, de façon assez surprenante il peut arriver qu'une statistique exhaustive minimale comprenne une statistique libre, qui intuitivement ne devrait pas être prise en compte pour donner toute l'information sur θ (soit par exemple la loi de IP_θ discrète et équidistribuée sur $\{\theta-1, \theta, \theta+1\}$; pour un échantillon de taille 2, la statistique $(X_{(2)} - X_{(1)}, X_1 + X_2)$ est exhaustive minimale, mais $X_{(2)} - X_{(1)}$ est libre). Aussi peut-on rajouter une autre caractérisation des statistiques exhaustives pour pouvoir atteindre une forme d'optimalité de ces statistiques, ce qui serait qu'aucune fonctionnelle non constante de la statistique ne soit libre. Cela peut également se traduire de la façon suivante:

Définition. Une statistique exhaustive T d'un modèle statistique paramétrique dominé avec T à valeur dans IR^d est dite complète si pour toute fonction borélienne $h : IR^d \to IR$ telle que $h(T)$ soit intégrable, alors:

$$\forall \theta \in \Theta, \quad IE_\theta\, h(T) = 0 \quad \Longrightarrow \quad h(T) = 0.$$

Propriété. Soit un modèle statistique paramétrique dominé.

1. si T est une statistique exhaustive complète alors pour toute fonction borélienne h bijective $h(T)$ est une statistique exhaustive complète.
2. si T est une statistique exhaustive complète alors T est une statistique exhaustive minimale.
3. (Théorème de Basu) Si T est une statistique exhaustive complète alors T est indépendante de toute statistique libre sur le modèle.

Démonstration de la propriété 3. Théorème de Basu. Soit S une statistique libre pour le modèle et soit f une fonction telle que $IE_\theta(f(S))$ existe. Comme S est libre, on peut noter $e(f) = IE_\theta(f(S))$ une application linéaire ne dépendant pas de θ. Par suite, la statistique $IE_\mu(f(S) \mid T) - e(f)$ est une fonction de T mesurable telle que $IE_\theta\left[IE_\mu(f(S) \mid T) - e(f)\right] = 0$ pour tout $\theta \in \Theta$. Comme on a supposé que T est exhaustive complète, alors $IE_\mu(f(S) \mid T) = e(f)$ presque-sûrement : les statistiques S et T sont indépendantes. ■

Définition. On suppose un modèle paramétrique $((\Omega')^n, \mathcal{A}'_n, IP_\theta, \theta \in \Theta \subset IR^p)$ dominé par une mesure μ. Si, pour tout $(x_1, \cdots, x_n) \in (\Omega')^n$ et $\theta \in \Theta$, la vraisemblance de ce modèle par rapport à μ peut s'écrire sous la forme:

$$L_\theta(x_1, \cdots, x_n) = \exp\left[\beta(\theta) + b(x_1, \cdots, x_n) + \sum_{j=1}^{p} a_j(x_1, \cdots, x_n) \cdot \alpha_j(\theta)\right], \qquad (3)$$

avec les fonctions $a_j : (\Omega')^n \to IR$, $b : (\Omega')^n \to IR$, $\alpha_j : \Theta \subset IR^p \to IR$, et $\beta : \Theta \to IR$, alors on dit que le modèle est exponentiel (ou qu'il appartient à la famille exponentielle).

Exemple. Appartiennent à la famille exponentielle les lois:

– Lois discrètes : Lois de Bernoulli, binomiales, de Poisson,...
– Loi "continues" : Lois normales, exponentielles, gamma, du chi-deux,...

Remarque. Si $(X_1, \cdots, X_n)$ est une statistique exhaustive du modèle exponentiel (avec Θ plus grand) alors l'ensemble des valeurs prises par $(X_1, \cdots, X_n)$ ne dépend pas du paramètre θ.

Propriété. Soit un modèle exponentiel. Si pour $\theta \in \Theta$ on note $\alpha(\theta) = (\alpha_1(\theta), \ldots, \alpha_p(\theta))$ et si l'ensemble $\alpha(\Theta)$ est d'intérieur non vide, alors $T = (x_1, \ldots, x_n) \to (a_1(x_1, \cdots, x_n), \ldots, a_p(x_1, \cdots, x_n))$ est une statistique exhaustive minimale et complète.

Démonstration de la propriété. Soit $g : \mathbb{R}^p \to \mathbb{R}$ telle que $\mathbb{E}_\theta(g(T)) = 0$. On a, $\forall \theta \in \Theta$,

$$\mathbb{E}_\theta(g(T)) = \int_{(\Omega')^n} g(T(x)) \cdot \exp\left(\beta(\theta) + b(x) + < T(x), \alpha(\theta) > \right) d\mu(x),$$

où $< ., . >$ désigne le produit scalaire. En remplaçant la mesure de densité $\exp(b(x))$ par rapport à μ, on obtient:

$$\mathbb{E}_\theta(g(T)) = 0 \Longrightarrow \int_{(\Omega')^n} g(T(x)) \cdot \exp(< T(x), \alpha(\theta) >) \, d\nu(x) = 0$$

$$\Longrightarrow \int_{T((\Omega')^n)} g(y) \exp(< y, \alpha(\theta) >) \, d\nu_T(y) = 0$$

pour tout $\theta \in \Theta$, en ayant noté ν_T la mesure image de ν par T et avec $T((\Omega')^n) \subset \mathbb{R}^p$. Si on note g^+ et g^- les parties positives et négatives de g (donc $g = g^+ - g^-$), et π^+ et π^- les mesures de densités g^+ et g^- par rapport à ν_T, on a alors, pour tout $\theta \in \Theta$:

$$\int_{T((\Omega')^n)} \exp(< y, \alpha(\theta) >) \, d\pi^+(y) = \int_{T((\Omega')^n)} \exp(< y, \alpha(\theta) >) \, d\pi^-(y).$$

En conséquence sur Θ, donc sur une partie d'intérieur non vide, les mesures π^+ et π^- ont des transformées de Laplace égales : ces deux mesures sont donc égales et donc $g^+ = g^-$ ν_T - presque partout (ce qui revient à $g = 0$). A partir des expressions des densités, on montre que $g = 0$, P_θ presque partout. ■

3.3 Information de Fisher

Pour mesurer l'information fournie par un échantillon (ou une statistique sur ce modèle) au sujet d'un paramètre, une idée naturelle serait de mesurer comment varie localement la mesure de probabilité, ou encore sa variance. Les fluctuations moyennes de cette variance semblance seraient donc un bon indicateur : pour ce faire on considère, lorsqu'il existe, $\frac{\partial}{\partial \theta} \log(L_\theta(X_1, \cdots, X_n))$, et on s'intéresse à la matrice de covariance de $\left(\frac{\partial}{\partial \theta} \log(L_\theta(X_1, \cdots, X_n))\right)$, dont on peut montrer qu'elle ne dépend pas du choix de la mesure dominante choisie. Précisons d'abord la notion de modèle régulier qui nous permettra de définir cette quantité d'information.

Définition. Dans le cadre d'un modèle statistique paramétrique $((\Omega', A'_n, P_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure μ, on dira que ce modèle est régulier lorsque:

1. Θ est un ouvert de $\mathbb{R}^p$
2. la vraisemblance (L_θ) vérifie $\forall (x_1, \ldots, x_n) \in (\Omega')^n$, $\forall \theta \in \Theta, L_\theta(x_1, \ldots, x_n) > 0$;
3. $\forall (x_1, \ldots, x_n) \in (\Omega')^n$, la fonction $\theta \in \Theta \to \log(L_\theta(.))$ est différentiable sur Θ par rapport à θ, et son gradient appartient à $\mathcal{L}^2((\Omega')^n, A'_n, P_\theta)$ $\forall \theta \in \Theta$;
4. $\forall \theta \in \Theta$, pour toute fonction $h : (\Omega')^n \to \mathbb{R}$ appartenant à $\mathcal{L}^1((\Omega')^n, A'_n, P_\theta)$, alors:

$$\frac{\partial}{\partial \theta} \int_{(\Omega')^n} h(x) \cdot L_\theta(x) \, d\mu(x) = \int_{(\Omega')^n} h(x) \cdot \frac{\partial}{\partial \theta} L_\theta(x) \, d\mu(x). \qquad (4)$$

Comme on le verra pour un modèle régulier, on a $\mathbb{E}_\theta(\text{grad}_\theta(\log L_\theta(.))) = 0$.

Démonstration : On a $\mathbb{E}_\mu(L_\theta(.)) = 1$ donc $\mathbb{E}_\mu(\text{grad}_\theta(L_\theta(.))) = 0$. Par conséquent $\mathbb{E}_\mu\left(\frac{\text{grad}_\theta(L_\theta(.))}{L_\theta(.)}\right) = 0$, soit $\mathbb{E}_\theta(\text{grad}_\theta(\log L_\theta(.))) = 0$. ■

Définition. Pour un modèle statistique paramétrique régulier, on appelle information de Fisher, la matrice:

$$I_n(\theta) = \mathbb{E}_\theta\left[\frac{\partial(\log L_\theta(X_1,\cdots,X_N))}{\partial\theta_i}\times\frac{\partial(\log L_\theta(X_1,\cdots,X_N))}{\partial\theta_j}\right]_{1\le i,j\le p}.$$

Propriété. Pour un modèle statistique paramétrique régulier, et $\forall (x_1,\ldots,x_n)\in(\Omega')^n$, la fonction $\theta\in\Theta\to\log(L_\theta(.))$ est $C^2(\Theta)$, alors:

$$I_n(\theta) = -\ \mathbb{E}_\theta\left[\frac{\partial^2(\log L_\theta(X_1,\cdots,X_N))}{\partial\theta_i\cdot\partial\theta_j}\right]_{1\le i,j\le p}.$$

Définition. L'information de Fisher $I_n^T(\theta)$ associée à une statistique T, lorsqu'elle existe, est la matrice de Fisher de la vraisemblance de T (définie à partir de la vraisemblance de T).

Propriété. Pour un modèle régulier, T est une statistique libre si et seulement si $I_n^T(\theta)=0$.

<u>Démonstration</u> : $\Rightarrow$ Si T est libre alors sa loi ne dépend pas de θ donc le gradient du logarithme de sa vraisemblance est nul ; l'information de Fisher associée à T est nulle.

$\Leftarrow$ Si $I_n^T(\theta)=0$, donc la statistique $\nabla_\theta(\log L_\theta^T(T))$ est centrée et de matrice de covariance nulle. Ainsi, pour tout $\theta\in\Theta$, il existe un ensemble N_θ de mesure 1 pour la mesure de probabilité définie par T (donc, d'après la première hypothèse d'un modèle régulier, tel que $\mu(N_\theta)=1$) et tel que pour tout $t\in N_\theta$, $\nabla_\theta(\log L_\theta^T(t))=0$. Pour montrer que $\nabla_\theta(\log L_\theta^T(t))=0$ est bien une variable aléatoire nulle μ-p.s., et donc que $\log L_\theta^T(.)$ est une fonction constante en θ, nous faut montrer que finalement les θ ne dépendent pas de θ. Soit $\Theta^{(d)}=\{\theta_i^{(d)}\}_{i\in\mathbb{N}}$ un sous-ensemble dénombrable dense dans Θ. Comme $\Theta^{(d)}$ est dénombrable, est clair que $N=\bigcap_{i\in\mathbb{N}}N_{\theta_i^{(d)}}$ est tel que $\mu(N)=1$. De plus, pour tout $\theta\in\Theta$, il existe une sous-suite $(\theta_{\varphi(n)}^{(d)})_{n\in\mathbb{N}}$ de $\Theta^{(d)}$ convergeant vers θ et telle que pour tout $t\in N$, pour tout $n\in\mathbb{N}$, $\nabla_{\theta_{\varphi(n)}^{(d)}}(\log L_{\theta_{\varphi(n)}^{(d)}}^T(t))=0$. Comme une telle fonction $\theta_{\varphi(n)}^{(d)}$ est continue, cette propriété passe à la limite, et donc pour tout $t\in N$, $\forall\theta\in\Theta$, $\nabla_\theta(\log L_\theta^T(t))=0$. Comme N ne dépend pas de θ, alors la fonction $\theta\to\log L_\theta^T(.)$ est une constante ne dépendant pas de θ, μ- p.s. : la statistique T est bien libre. ■

Propriété. Pour un modèle régulier, si T est une statistique exhaustive, $I_n^T(\theta)=I_n(\theta)$ pour tout $\theta\in\Theta$.

<u>Démonstration</u> : Comme T est une statistique exhaustive, on peut écrire d'après la démonstration du Théorème de factorisation de Neyman que pour tout $(x_1,\ldots,x_n)\in(\Omega')^n$ et tout $\theta\in\Theta$:

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}^*}(x_1,\ldots,x_n)=g_\theta(T(x_1,\ldots,x_n)).$$

On peut écrire cela pour la mesure image : $\frac{d\mathbb{P}_\theta^T}{d\mathbb{P}^{*T}}(t)=g_\theta(t)$, pour tout $t\in T((\Omega')^n)$ et tout $\theta\in\Theta$. En conséquence, pour tout $\theta\in\Theta$,

$$\begin{aligned}
I(\theta) &= \mathbb{E}_\theta\left[\frac{\partial(\log L_\theta(X_1,\cdots,X_N))}{\partial\theta_i}\times\frac{\partial(\log L_\theta(X_1,\cdots,X_N))}{\partial\theta_j}\right]_{1\le i,j\le p}\\
&= \int_{(\Omega')^n}\left[\frac{\partial(\log L_\theta(x))}{\partial\theta_i}\times\frac{\partial(\log L_\theta(x))}{\partial\theta_j}\right]_{1\le i,j\le p}d\mathbb{P}_\theta(x)\\
&= \int_{(\Omega')^n}\left[\frac{\partial(\log g_\theta(T(x)))}{\partial\theta_i}\times\frac{\partial(\log g_\theta(T(x)))}{\partial\theta_j}\right]_{1\le i,j\le p}g_\theta(T(x))d\mathbb{P}^*(x) \quad \text{car } \log L_\theta(x)=\log g_\theta(T(x))+\log h(x)\\
&= \int_{T(\Omega')^n}\left[\frac{\partial(\log g_\theta(t))}{\partial\theta_i}\times\frac{\partial(\log g_\theta(t))}{\partial\theta_j}\right]_{1\le i,j\le p}g_\theta(t)d\mathbb{P}^{*T}(x) \quad \text{d'après le théorème du transport}\\
&= \int_{T(\Omega')^n}\left[\frac{\partial(\log g_\theta(t))}{\partial\theta_i}\times\frac{\partial(\log g_\theta(t))}{\partial\theta_j}\right]_{1\le i,j\le p}d\mathbb{P}_\theta(t)\\
&= I_n^T(\theta).
\end{aligned}$$

■

Remarque. En rajoutant certaines hypothèses de continuité sur la vraisemblance de T, on peut montrer que la réciproque est également vraie, et donc que $I_n^T(\theta) = 0$ si et seulement si la statistique T est exhaustive.

Ainsi, on retrouve à l'aide de la notion d'information de Fisher les "intuitions" qui ont guidé la section précédente. Voyons maintenant les applications de la notion d'exhaustivité à l'estimation paramétrique.

3.4 Application à l'estimation paramétrique

On se place dans le cadre d'un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure μ. Par ailleurs, on suppose que Θ est un ouvert.

Définition.
– Soit $g : \Theta \to \Theta'$, où $\Theta' \subset \mathbb{R}^{p'}$ avec $p' \in \mathbb{N}^*$, une fonction mesurable. On appelle estimateur de la fonction g du paramètre, donc de $g(\theta)$, une statistique T à valeurs dans $\mathbb{R}^{p'}$. En particulier, un estimateur du paramètre θ est une statistique à valeurs dans $\mathbb{R}^p$. Une estimation de $g(\theta)$ est une réalisation de T.
– On appelle le biais d'un estimateur T de $g(\theta)$ le vecteur constant de $\mathbb{R}^{p'}$, $B(\theta) = \mathbb{E}_\theta(T) - g(\theta)$. On dira que l'estimateur est sans biais si $B(\theta) = 0$ pour tout $\theta \in \Theta$.
– On appelle le risque quadratique de l'estimateur T de $g(\theta)$ le réel positif $R(\theta) = \mathbb{E}_\theta(\|T - g(\theta)\|^2)$, où $\|.\|$ désigne usuellement la norme euclidienne (mais cela peut être une autre fonctionnelle positive et convexe). Si l'estimateur est sans biais alors, $R(\theta) = \mathrm{Trace}(\mathrm{cov}(T))$.

Pour pouvoir parler du comportement asymptotique d'une statistique, on va devoir se placer dans un modèle dans lequel un échantillon est une suite de v.a.; En quelque sorte ce gros modèle pourra s'écrire $((\Omega')^{\mathbb{N}}, \mathcal{A}'_{\mathbb{N}}, \mathbb{P}_\theta^{\mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$ (la dimension du paramètre reste constante). Pour un n fixé, une statistique T_n sera d'abord une projection du "gros" modèle sur le modèle de taille n, puis une statistique "normale". On devra donc parler d'une suite d'estimateurs $(T_n)_n$.

Définition. Pour un modèle statistique paramétrique $((\Omega')^{\mathbb{N}}, \mathcal{A}'_{\mathbb{N}}, \mathbb{P}_\theta^{\mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, et pour $(T_n)_n$ une suite d'estimateurs de $g(\theta)$:
– Si $\lim_{n\to\infty} B_n(\theta) = 0$, on dit que l'estimateur est asymptotiquement sans biais.
– On dit que $(T_n)_n$ est convergent lorsque $T_n \xrightarrow[n\to+\infty]{\mathbb{P}} g(\theta)$.
– S'il existe (a_n) une suite de réels positifs tels que $a_n(T_n - g(\theta)) \xrightarrow[n\to+\infty]{\mathcal{L}} Z_\theta$, où Z_θ est une loi centrée non nulle (ne dépendant pas de n), on dit (T_n) converge vers $g(\theta)$ à la vitesse a_n.

A priori, être sans biais n'est pas un critère pour garantir une certaine optimalité de convergence d'un estimateur. On préfère alors plutôt discriminer entre deux potentiels estimateurs à l'aide d'un critère sur le risque quadratique ou sur la matrice de variance-covariance. Cependant, il n'existe pas de résultats généraux pour trouver un "meilleur" estimateur en ce sens. Pour en obtenir, on devra se limiter à une certaine classe d'estimateurs, celle des estimateurs sans biais.

Définition. Soit un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, et soit T un estimateur sans biais de $g(\theta)$. On dit que T est de variance uniformément minimum parmi les estimateurs sans biais de $g(\theta)$ lorsque pour tout estimateur sans biais de $g(\theta)$ S, on a, $\forall \theta \in \Theta$, $\mathrm{cov}(T) \leq \mathrm{cov}(S)$ (au sens où $\mathrm{cov}(T) - \mathrm{cov}(S)$ est une matrice positive).

Propriété. Si T est un estimateur de variance uniformément minimum parmi les estimateurs sans biais, alors il est unique $\mathbb{P}_\theta$-p.s.

Démonstration : Soit S un autre estimateur que l'on suppose également de variance uniformément minimum parmi les estimateurs sans biais. Montrons d'abord que $\mathbb{E}_\theta((T - S) \cdot {}^tT) = 0$. En effet, si $\alpha \in \mathbb{R}$, comme T est de variance minimum, en utilisant l'ordre défini sur les matrices symétriques :

$$
\begin{aligned}
\mathrm{cov}(T) &\leq \mathrm{cov}(T + \alpha(T - S)) \\
&\leq \mathrm{cov}(T) + \alpha^2 \mathrm{cov}(T - S) + 2\alpha \cdot \mathbb{E}_\theta(T \cdot {}^tS) \\
\Longrightarrow \quad 0 &\leq \alpha \cdot \alpha \cdot \mathrm{cov}(S) + 2\alpha\, \mathbb{E}_\theta(T \cdot {}^t(T - S)) \qquad \text{pour tout } \alpha \in \mathbb{R}.
\end{aligned}
$$

Comme $\mathrm{cov}_\theta(T-S)$ est une matrice positive, il existe une matrice A telle que $\mathrm{cov}_\theta(T-S) = A \cdot {}^t A$. Il vient alors: $\mathrm{IE}_\theta(T \cdot {}^t(T-S)) = 0$. Par suite, comme $\mathrm{cov}_\theta(T-S) = \mathrm{IE}_\theta\,(T-S)\cdot{}^t(T-S) = \mathrm{IE}_\theta\, T\cdot{}^t(T-S) - \mathrm{IE}_\theta\, S\cdot{}^t(T-S)$, et que l'on a supposé T et S de variance minimale, $\mathrm{cov}_\theta(T-S) = 0$. Donc $T = S$ sur un ensemble de IP_θ-mesure égale à 1. ■

Théorème (Rao-Blackwell): Si T est un estimateur sans biais de $g(\theta)$ et si S est une statistique exhaustive, alors $R = \mathrm{IE}_\theta(T \mid S)$, qui ne dépend pas de θ car S est exhaustive, est un estimateur sans biais de $g(\theta)$ de matrice de covariance inférieure ou égale à celle de T.

Démonstration: On a déjà vu que $\mathrm{IE}_\theta(R) = \mathrm{IE}_\theta(T) = g(\theta)$. De plus, pour tout $u \in \mathrm{IR}^{p'}$ (avec $g : \mathrm{IR}^p \to \mathrm{IR}^{p'}$),

$$\begin{aligned}
\mathrm{cov}_\theta(u \cdot T) &= \mathrm{IE}_\theta\ {}^t u \cdot (T - g(\theta))^2 \\
&= \mathrm{IE}_\theta\ \mathrm{IE}_\theta\ {}^t u \cdot (T - g(\theta))^2 \mid S \\
&\geq \mathrm{IE}_\theta\ \mathrm{IE}_\theta\ {}^t u \cdot (T - g(\theta)) \mid S^2 \quad \text{d'après l'inégalité de Jensen,} \\
&\geq \mathrm{cov}_\theta(u \cdot R).
\end{aligned}$$

Ceci a bien entraîné que $\mathrm{cov}(T) \geq \mathrm{cov}(R)$. ■

Théorème (Lehmann-Scheffé): Si T est un estimateur sans biais de $g(\theta)$ et si S est une statistique exhaustive et complète, alors l'unique estimateur de $g(\theta)$ sans biais uniformément de variance minimale est $R = \mathrm{IE}_\theta(T \mid S)$ (c'est-à-dire que R est une fonction de S).

Démonstration: Soit T' un autre estimateur sans biais de $g(\theta)$. Si $R' = \mathrm{IE}_\theta(T' \mid S)$, on sait que $\mathrm{cov}(T') \geq \mathrm{cov}(R')$ d'après le Théorème de Rao-Blackwell. Or $\mathrm{IE}_\theta(R - R') = 0$ pour tout $\theta \in \Theta$ car les deux estimateurs sont sans biais. De plus comme R et R' sont des fonctions de S, $R - R'$ l'est aussi, et du fait que S est une statistique exhaustive et complète, pour tout $\theta \in \Theta$, $R = R'$, IP_θ-p.s. Par conséquent, pour tout $\theta \in \Theta$, $\mathrm{cov}(R') = \mathrm{cov}(R)$ et donc $\mathrm{cov}(R) = \mathrm{cov}(T')$: R est bien l'estimateur sans biais de variance uniformément minimale. ■

Retenons donc de tout ceci que l'estimateur sans biais de $g(\theta)$ et de variance uniformément minimale est une unique fonction d'une statistique exhaustive et complète, lorsqu'une telle statistique existe. On aimerait maintenant connaître un peu mieux la covariance d'un tel estimateur.

Théorème (Inégalité de Cramer-Rao): Soit un modèle statistique paramétré $((\Omega^n, A_n, \mathrm{IP}_\theta, \theta \in \Theta)$ dominé et régulier, et soit T un estimateur sans biais de $g(\theta)$, tel que $\mathrm{IE}_\theta T^2 < +\infty$. Si on suppose que l'information de Fisher est une matrice définie positive, alors, en notant $\frac{\partial g}{\partial \theta}(\theta)$ la matrice jacobienne de g, pour tout $\theta \in \Theta$:

$$\mathrm{cov}(T) \geq \frac{\partial g}{\partial \theta}(\theta) \cdot (I_n(\theta))^{-1} \cdot {}^t\frac{\partial g}{\partial \theta}(\theta) \quad \text{(au sens des matrices symétriques).}$$

En particulier, si T est un estimateur sans biais de θ, alors:

$$\mathrm{cov}(T) \geq (I_n(\theta))^{-1} \quad \text{(au sens des matrices symétriques).}$$

Démonstration: Soit $Z_\theta(x) = \mathrm{grad}(\log L_\theta(x))$ où $x \in (\Omega')^n$ suit IP_θ. On sait que comme le modèle est régulier, $\mathrm{IE}_\theta(Z_\theta) = 0$ pour tout $\theta \in \Theta$ et donc:

$$\mathrm{cov}(Z_\theta) = I_n(\theta) \quad \text{pour tout } \theta \in \Theta.$$

De plus, T est un estimateur sans biais de $g(\theta)$ donc pour tout $\theta \in \Theta$

$$\begin{aligned}
\mathrm{IE}_\theta(T) = g(\theta) &\Longrightarrow \int_{(\Omega')^n} T(x) \cdot \frac{\partial L_\theta}{\partial \theta}(x)\, d\mu(x) = \frac{\partial g}{\partial \theta}(\theta) \quad \text{(en dérivant)} \\
&\Longrightarrow \int_{(\Omega')^n} T(x) \cdot \frac{\partial L_\theta}{\partial \theta}(x) \cdot (L_\theta(x))^{-1}\, d\mathrm{IP}_\theta(x) = \frac{\partial g}{\partial \theta}(\theta) \\
&\Longrightarrow \mathrm{IE}_\theta(T \cdot {}^t Z_\theta) = \frac{\partial g}{\partial \theta}(\theta).
\end{aligned}$$

Ainsi, d'après ce qui précède,

$$\begin{aligned} \mathrm{cov}_\theta\Big(T - \frac{\partial g}{\partial \theta}(\theta)\cdot I^{-1}(\theta)\cdot Z_\theta\Big) &= \mathrm{cov}_\theta(T) - 2\frac{\partial g}{\partial \theta}(\theta)\cdot I^{-1}(\theta)\cdot {}^t\frac{\partial g}{\partial \theta}(\theta) + \frac{\partial g}{\partial \theta}(\theta)\cdot I^{-1}(\theta)\cdot {}^t\frac{\partial g}{\partial \theta}(\theta) \\ &= \mathrm{cov}_\theta(T) - \frac{\partial g}{\partial \theta}(\theta)\cdot I^{-1}(\theta)\cdot {}^t\frac{\partial g}{\partial \theta}(\theta). \end{aligned}$$

En conséquence, comme $\mathrm{cov}_\theta\big(T - \frac{\partial g}{\partial \theta}(\theta)\cdot I^{-1}(\theta)\cdot Z_\theta\big)$ est une matrice positive, l'inégalité de Cramer-Rao est prouvée. ■

Corollaire. Deux cas particuliers méritent attention:

– Si le modèle est de la forme $((\Omega')^n, \mathcal{A}'_n, (f_\theta\cdot d\mu)^{\otimes n}, \theta\in\Theta)$, alors $I_n(\theta) = n\cdot I_1(\theta)$, où $I_1(\theta)$ est la matrice d'information de Fisher d'une seule variable aléatoire X distribuée suivant $f_\theta \cdot d\mu$ et l'inégalité de Cramer-Rao devient donc:

$$\mathrm{cov}(T) \geq \frac{1}{n}\cdot \frac{\partial g}{\partial \theta}(\theta)\cdot (I_1(\theta))^{-1}\cdot {}^t\frac{\partial g}{\partial \theta}(\theta) \qquad \text{(au sens des matrices symétriques).}$$

On voit donc que pour un échantillon de variables indépendantes et identiquement distribuées, si la vraisemblance est régulière, alors la vitesse de convergence de tout estimateur sans biais est au mieux en $\sqrt{n}$.

– Si le modèle n'est pas régulier, mais que sous la probabilité IP_θ, la matrice d'information de Fisher existe et est inversible, surtout si la propriété (4) est vérifiée, alors l'inégalité de Cramer-Rao est vérifiée. Cela exclut cependant les modèles dont le support de IP_θ dépend de θ, comme par exemple le simple modèle de v.a.i.i.d. de loi $\mathcal{U}(]0, \theta[)$, avec $\theta > 0$.

Définition. Si un estimateur sans biais atteint (respectivement asymptotiquement) la borne de Cramer-Rao (qui ne dépend pas de l'estimateur), on dit qu'il est (resp. asymptotiquement) efficace.

Remarque. Un estimateur peut être sans biais, de variance minimale, mais ne pas atteindre la borne de Cramer-Rao, donc ne pas être efficace. De la même manière, il peut exister des estimateurs biaisés atteignant la borne de Cramer-Rao.

Nous allons voir que les modèles exponentiels jouent un rôle central pour l'estimation puisque ce sont sous certaines conditions ils sont les seuls pour lesquels on aura une estimation sans biais efficace.

Théorème. Soit un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, IP_\theta, \theta\in\Theta)$, avec $\Theta \subset IR^p$, dominé et régulier. Soit $g : IR^d \to IR^d$ de classe C^1 sur Θ telle que la matrice carrée de taille p, $\frac{\partial g}{\partial \theta}(\theta)$ soit de rang p pour tout $\theta\in\Theta$. Alors $T = {}^t(T_1, \ldots, T_d)$ est un estimateur sans biais de $g(\theta)$ atteignant la borne de Cramer-Rao si et seulement si le modèle est exponentiel et plus précisément s'il existe des fonctions $a : (\Omega')^n \to IR$, $\beta : \Theta \to IR$ et $\alpha_j : \Theta \to IR$ $(1 \leq j \leq p)$, telles que pour tout $\theta\in\Theta$, $g(\theta) = -\Big(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\Big)^{-1}_{1\leq i,j\leq p}\cdot \frac{\partial \beta}{\partial \theta}(\theta)$ et

$$L_\theta(x_1,\cdots,x_n) = \exp\Big(\beta(\theta) + b(x_1,\cdots,x_n) + \sum_{j=1}^{d} T_j(x_1,\cdots,x_n)\cdot \alpha_j(\theta)\Big).$$

Démonstration : ⇐= On suppose donc le modèle exponentiel décrit dans le théorème. Si on dérive par rapport à θ un tel modèle, on obtient que pour μ-presque tout $x\in(\Omega')^n$

$$\frac{\partial}{\partial \theta}(\log L_\theta(x)) = \Big(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\Big)_{1\leq i,j\leq p}\cdot T + \frac{\partial \beta}{\partial \theta}(\theta), \quad \text{pour tout } \theta\in\Theta. \tag{5}$$

En conséquence, comme $I(\theta) = IE_\theta\Big[\frac{\partial}{\partial \theta}(\log L_\theta(.))\cdot {}^t\frac{\partial}{\partial \theta}(\log L_\theta(.))\Big]$, on en déduit que:

$$I(\theta) = \Big(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\Big)_{1\leq i,j\leq d}\cdot \mathrm{cov}_\theta(T)\cdot {}^t\Big(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\Big)_{1\leq i,j\leq p} \Longrightarrow \mathrm{cov}_\theta(T) = \Big(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\Big)^{-1}_{1\leq i,j\leq p}\cdot I(\theta)\cdot {}^t\Big(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\Big)^{-1}_{1\leq i,j\leq p}$$

Par ailleurs, comme T est un estimateur sans biais de g (θ), la preuve de l'inégalité de Cramer - Rao,

$$\mathbb{E}_\theta\left[T(.)\cdot {}^t\frac{\partial}{\partial\theta}(\log L_\theta(.))\right] = \frac{\partial g}{\partial\theta}(\theta)$$

et en utilisant (5) que l'on multiplie à gauche par $\frac{\partial}{\partial\theta}(\log L_\theta(.))$, on obtient:

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}(\log L_\theta(.))\cdot {}^t\frac{\partial}{\partial\theta}(\log L_\theta(.))\right] = \mathbb{E}_\theta\left[\frac{\partial\alpha_j}{\partial\theta_i}(\theta)_{1\le i,j\le p}\cdot T\cdot {}^t\frac{\partial}{\partial\theta}(\log L_\theta(.))\right] + \mathbb{E}_\theta\left[\frac{\partial\beta}{\partial\theta}(\theta)\cdot {}^t\frac{\partial}{\partial\theta}(\log L_\theta(.))\right],$$

et donc $I(\theta) = \frac{\partial\alpha_j}{\partial\theta_i}(\theta)_{1\le i,j\le p}\cdot\frac{\partial g}{\partial\theta}(\theta)$. A l'aide de cette égalité, en remplaçant le calcul précédent, on en arrive à ce que:

$$\mathrm{cov}_\theta(T) = \frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot {}^t\frac{\partial g}{\partial\theta}(\theta),$$

donc T atteint bien la borne de Cramer-Rao. De plus, grâce à (5),

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}(\log L_\theta(x))\right] = \mathbb{E}_\theta\left[\frac{\partial\alpha_j}{\partial\theta_i}(\theta)_{1\le i,j\le p}\cdot T\right] + \frac{\partial\beta}{\partial\theta}(\theta)$$

$$\text{soit} \qquad 0 = \frac{\partial\alpha_j}{\partial\theta_i}(\theta)_{1\le i,j\le p}\cdot g(\theta) + \frac{\partial\beta}{\partial\theta}(\theta)$$

$$\text{et donc} \qquad g(\theta) = -\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)_{1\le i,j\le p}\right)^{-1}\cdot\frac{\partial\beta}{\partial\theta}(\theta).$$

$\Longrightarrow$ D'après la preuve de l'égalité de Cramer-Rao, si T est un estimateur sans biais de $g(\theta)$ atteignant la borne de Cramer-Rao, alors

$$\mathrm{cov}_\theta\left(T - \frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot Z_\theta\right) = 0.$$

Ainsi, pour tout $\theta\in\Theta$, il existe un ensemble $N_\theta\subset(\Omega')^n$ tel que $P_\theta(N_\theta) = 1$ et tel que pour tout $x\in N_\theta$, $T(x) - g(\theta) = \frac{\partial g}{\partial\theta}(\theta)\cdot I^{-1}(\theta)\cdot Z_\theta(x)$. Par le même procédé que celui de la preuve de la borne de l'information de Fisher pour une statistique libre, on peut trouver un ensemble N ne dépendant pas de θ, tel que cette propriété soit également vraie, avec $\mu(N) = 1$, ce qui revient à poser $N = \bigcap_{\theta\in\Theta}N_\theta$,

$$I(\theta)\cdot\left(\frac{\partial g}{\partial\theta}(\theta)\right)^{-1}\cdot\left(T(x) - g(\theta)\right) = \frac{\partial}{\partial\theta}(\log L_\theta(x)), \quad \text{pour tout } \theta\in\Theta.$$

Alors en intégrant par rapport à θ, en notant

- $\alpha(\theta)$ le vecteur colonne primitive de $I(\theta)\cdot\left(\frac{\partial g}{\partial\theta}(\theta)\right)^{-1}$
- $\beta(\theta)$ la fonction primitive de $-\alpha(\theta)\cdot\left(\frac{\partial g}{\partial\theta}(\theta)\right)^{-1}\cdot g(\theta)$
- $b(x)$ une fonction ne dépendant pas de θ

on a $\log L_\theta(x) = \alpha(\theta)\cdot T(x) + \beta(\theta) + b(x)$, d'où la structure de la vraisemblance sous forme de l'exponentielle, et on retrouve l'expression de $g(\theta)$ par le raisonnement que plus haut. ■

Corollaire: A l'inverse, si l'on dispose d'un modèle exponentiel régulier (3), alors il existe une fonction (à une translation affine près) du paramètre pour laquelle l'estimateur est efficace: il s'agit de $g(\theta) = -\frac{1}{n}\cdot\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)_{1\le i,j\le p}\right)^{-1}\cdot\frac{\partial\beta}{\partial\theta}(\theta)$ (noter que cette fonction semble dépendre de n ; dans le cas de v.a.i.i.d. ce n'est pas le cas). L'estimateur est alors $\frac{1}{n}\cdot T = \frac{1}{n}(\alpha(X_1,\ldots,X_n),\ldots,\alpha(X_1,\ldots,X_n))$ et sa matrice de covariance minimale est donnée par sa borne de Cramer- Rao, soit:

$$\mathrm{cov}_\theta(T) = \frac{1}{n}\cdot\frac{\partial g}{\partial\theta}(\theta)\cdot\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\le i,j\le d}.$$

3.5 Estimateur du maximum de vraisemblance

Nous allons voir une méthode permettant d'obtenir ainsi et dans la plupart des cas un estimateur possédant de bonnes qualités. Par la suite on se place une nouvelle fois dans le cadre d'une méthode statistique paramétrique $((\Omega', A', I\!P_\theta, \theta \in \Theta)$, avec $\Theta \subset I\!R^p$, domaine.

Définition. Pour $(x_1, \ldots, x_n) \in (\Omega')^n$, soit $\theta \in \Theta \to L_\theta(x_1, \ldots, x_n)$ la vraisemblance du modèle. On appelle estimateur du maximum de vraisemblance un estimateur θ_n tel que pour $(X_1, \ldots, X_n)$ un n-échantillon quelconque du modèle

$$L_{\theta_n}(X_1, \cdots, X_n) = \sup_{\theta \in \Theta} L_\theta(X_1, \cdots, X_n).$$

Remarque. Il n'y a pas de garantie de l'unicité d'un tel estimateur. Une méthode pour l'obtenir (mais pas toujours) est de rechercher un extremum local de L_θ sur Θ, ce qui pourra être fait en annulant les dérivées partielles de L_θ par θ_i. De même, il est clair que l'estimateur du maximum de vraisemblance pourra être également obtenu en maximisant le logarithme de la vraisemblance, appelé log-vraisemblance.
En fin, si l'on désire estimer $g(\theta)$ avec g une fonction bijective, alors $g(\theta)$ sera l'estimateur du maximum de vraisemblance de $g(\theta)$.

Propriété. S'il existe une statistique exhaustive T pour le modèle, alors le maximum de vraisemblance est une fonction mesurable de T pour tout $\theta \in \Theta$.

Démonstration: Si T est exhaustive, d'après le théorème de factorisation, la vraisemblance du modèle par rapport à la mesure dominante est $g_\theta(T(x_1, \ldots, x_n))$ pour tout $\theta \in \Theta$ et $I\!P_\theta$- presque tout $(x, x_n) \in (\Omega')^n$, ce qui revient à dire $I\!P_\theta$-presque tout $(x, x_n) \in (\Omega')^n$ par la même démonstration que celle de la nullité de l'information de Fisher d'une statistique exhaustive. Ainsi, pour trouver le maximum de L_θ revient à prendre l'argument maximal de $g_\theta(T(x_1, \ldots, x_n))$, et θ sera donc une fonction de T. ■

Propriété. On suppose que le modèle est régulier. Si on suppose qu'il existe un estimateur sans biais efficace de θ alors c'est l'estimateur du maximum de vraisemblance de θ.

Démonstration: D'après ce qui précède, si le modèle est régulier et que T est un estimateur sans biais efficace de θ, alors le modèle est exponentiel (1.5) a encore lieu, soit pour tout $\theta \in \Theta$,

$$\frac{\partial}{\partial\theta}(\log L_\theta(x)) = \left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1 \le i, j \le p} \cdot T + \frac{\partial \beta}{\partial \theta}(\theta) \Longrightarrow \left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1 \le i, j \le p} \cdot I\!E_\theta(T) + \frac{\partial \beta}{\partial \theta}(\theta) = 0.$$

Comme T est un estimateur sans biais de θ, on a donc $\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1 \le i, j \le p} \cdot \theta + \frac{\partial \beta}{\partial \theta}(\theta) = 0$, pour tout $\theta \in \Theta$,

ce qui s'applique également à θ et donc:

$$\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1 \le i, j \le p} \cdot \theta + \frac{\partial \beta}{\partial \theta}(\theta) = 0.$$

Mais d'après sa définition, le modèle étant régulier θ minimise la log-vraisemblance et annule donc sa dérivée, ce qui implique que:

$$\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1 \le i, j \le p} \cdot T + \frac{\partial \beta}{\partial \theta}(\theta) = 0.$$

En conséquence, obtient:

$$\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1 \le i, j \le p} \cdot (T - \theta) = 0 \Rightarrow T = \theta,$$

car la matrice des dérivées est supposée de rang d. Enfin, l'unicité étant liée à l'écriture du modèle exponentiel.
■
Nous allons nous intéresser maintenant au comportement asymptotique de l'estimateur du maximum de vraisemblance (lorsqu'il existe), donc quand la taille de l'échantillon tend vers l'infini. Il est clair que pour chaque réalisation de l'expérience, toutefois, le modèle statistique change. Pour pallier à cela, on se placera dans un "grossissement" $(\Omega^{IN}, A_{IN}, I\!P_\theta^{IN}, \theta \in \Theta)$, où $\Theta \subset I\!R^p$ (la dimension du paramètre reste constante) dans lequel l'échantillon est une suite de v.a.. On supposera également que tout échantillon de ce modèle est constitué de v. a. i. i. d. , et que $I\!P_\theta^{IN} = (f_\theta \cdot d\mu)^{\otimes IN}$, le modèle étant dominé par la mesure μ, f_θ étant la densité chaque (X_i) par rapport à μ.

Théorème (Convergence de l'estimateur du maximum de vraisemblance). On suppose que le modèle $((\Omega')^{\mathbb{N}}, A'^{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^d$ domine par une mesure quelconque. On suppose en plus que le modèle est identifiable (au sens où $\theta_1 \neq \theta_2 \Rightarrow f_{\theta_1} \neq f_{\theta_2}$ μ-presque partout, ou encore $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$). Alors si la suite $(X_n)_{n \in \mathbb{N}}$ est issue du modèle avec pour paramètre $\theta_0 \in \Theta$,

$$\theta_n \xrightarrow[n \to +\infty]{p.s.} \theta_0 \quad \text{pour la mesure } (f_{\theta_0} d\mu)^{\otimes \mathbb{N}}.$$

<u>Démonstration</u> : En premier lieu, pour un θ fixé, il est clair que pour tout $i \in \mathbb{N}$,

$$\log(L_\theta(x_1, \ldots, x_n)) - \log(L_{\theta_0}(x_1, \ldots, x_n)) = \sum_{i=1}^{n} \log \frac{f_\theta(x_i)}{f_{\theta_0}(x_i)} .$$

Par ailleurs, pour tout $i \in \mathbb{N}$, les X_i ont toutes la même loi et pour tout $\theta \in \Theta$,

$$\begin{aligned} \mathbb{E}_{\theta_0} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} &\leq \log \mathbb{E}_{\theta_0} \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \qquad \text{(Inégalité de Jensen pour la fonction log)} \\ &\leq \log(\mathbb{E}[f_\theta(X_i)]) \\ &\leq 0. \end{aligned}$$

En fait, du fait que la fonction log est strictement concave, la borne 0 ne peut être atteinte que si $\theta = \theta_0$. Ainsi, avec la contrainte d'un modèle identifiable, si $\theta \neq \theta_0$ alors :

$$\mathbb{E}_{\theta_0} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0.$$

On peut appliquer la loi forte des grands nombres pour les variables $\log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}$ $_{i \in \mathbb{N}}$ (qui sont bien i.i.d. et L^1 car le modèle est régulier), et ainsi :

$$\frac{1}{n}(\log(L_\theta(X_1, \ldots, X_n)) - \log(L_{\theta_0}(X_1, \ldots, X_n))) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}$$
$$\xrightarrow[n \to +\infty]{p.s.} \mathbb{E}_{\theta_0} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0,$$

la convergence presque sûre ayant lieu pour la mesure $(f_{\theta_0} d\mu)^{\otimes \mathbb{N}}$. Considérons maintenant $\varepsilon > 0$ une famille dénombrable $(\theta_i^{(\varepsilon)})_{i \in I}$ dense sur la sphère de centre θ_0 et de rayon ε. Du fait du caractère dénombrable de cette famille, pour tout $\varepsilon > 0$, il existe pour tout $i \in I$, pour tout ε :

$$\log(L_{\theta_i^{(\varepsilon)}}(X_1, \ldots, X_n)) < \log(L_{\theta_0}(X_1, \ldots, X_n)) \quad \text{p.s. pour la mesure } (f_{\theta_0} d\mu)^{\otimes \mathbb{N}}.$$

Comme le modèle est régulier, pour tout $n \in \mathbb{N}^*$, la log-vraisemblance de $X_1, \ldots, X_n$ est continue sur Θ. De plus pour tout n elle atteint son unique maximum en θ_n, donc θ_n se trouve à l'intérieur de la boule de centre θ_0 de rayon ε (toujours p.s. pour la mesure $(f_{\theta_0} d\mu)^{\otimes \mathbb{N}}$). Le raisonnement étant vrai pour tout $\varepsilon > 0$, le théorème s'en déduit. ■

Théorème (Normalité asymptotique de l'estimateur du maximum de vraisemblance). On suppose que le modèle paramétrique $((\Omega')^{\mathbb{N}}, A'^{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, domine par une mesure quelconque. On suppose en plus que le modèle est identifiable et que la fonction $\theta \mapsto L_\theta$ est de classe $C^2(\Theta)$. Alors si la suite $(X_n)_{n \in \mathbb{N}}$ est issue du modèle avec pour paramètre $\theta_0 \in \Theta$:

$$\sqrt{n}(\theta_n - \theta_0) \xrightarrow[n \to +\infty]{L} N_d(0, I_1^{-1}(\theta_0)),$$

où $I_1(\theta)$ est la matrice de Fisher de taille p (supposée inversible) pour la variable X_1.

<u>Démonstration</u> : Comme le modèle est régulier, on peut différencier la vraisemblance et pour tout θ, noter :

$$M_\theta(X_1, \ldots, X_n) = \frac{1}{n} \frac{\partial}{\partial \theta} \log L_\theta(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log(f_\theta(X_i)).$$

Un développement limité d'ordre 1 de $\theta \mapsto M_\theta$ autour de θ_0 est possible (toujours en raison du théorème des accroissements finis) et donc pour tout $\theta \in \Theta$

$$M_\theta(X_1, \ldots, X_n) = M_{\theta_0}(X_1, \ldots, X_n) + (\theta - \theta_0) \cdot \frac{\partial}{\partial \theta} M_{\theta^*}(X_1, \ldots, X_n),$$

avec θ^* dans le segment $[\theta_0, \theta]$ (remarquons que $\frac{\partial}{\partial \theta} M_{\theta^*}(X_1, \ldots, X_n)$ est une matrice carrée de taille d). Ainsi en remplaçant θ par θ_n, on obtient pour chaque n l'existence d'un θ_n^* appartenant au segment $[\theta_0, \theta_n]$ tel que:

$$M_{\theta_n}(X_1, \ldots, X_n) = M_{\theta_0}(X_1, \ldots, X_n) + (\theta_n - \theta_0) \cdot \frac{\partial}{\partial \theta} M_{\theta_n^*}(X_1, \ldots, X_n). \tag{6}$$

Pour un modèle régulier, on a vu que $I_{\theta_0} = -\mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta_0}(X_i)\right] = -I_1(\theta_0)$, matrice de Fisher pour n'importe quelle variable X_i. Ainsi, $\frac{\partial}{\partial \theta} M_\theta(.)$ étant une moyenne empirique, on a par la loi forte des grands nombres:

$$\frac{\partial}{\partial \theta} M_{\theta_0}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_{\theta_0}(X_i) \xrightarrow[n \to +\infty]{\text{p. s.}} -I_1(\theta_0) \text{ pour la mesure } (f_{\theta_0} \cdot \mu)^{\otimes \mathbb{N}}.$$

Maintenant, en utilisant le fait que les θ_n^* sont dans Θ (classe C^2) et en utilisant la convergence presque sûre de θ_n vers θ_0 démontrée au théorème précédent, on a:

$$\frac{\partial}{\partial \theta} M_{\theta_n^*}(X_1, \ldots, X_n) \xrightarrow[n \to +\infty]{\text{p. s.}} -I_1(\theta_0) \text{ pour la mesure } (f_{\theta_0} \cdot \mu)^{\otimes \mathbb{N}}.$$

Finalement, comme θ_n est le maximum d'une fonction de classe C^1, cet estimateur annule $M_{\theta_n}(X_1, \ldots, X_n)$, et donc l'égalité (6) devient:

$$M_{\theta_0}(X_1, \ldots, X_n) \cdot I_1^{-1}(\theta_0) = (\theta_n - \theta_0).$$

Enfin, comme $M_{\theta_0}(X_1, \ldots, X_n)$ est une moyenne empirique avec terme à l'origine vérifie un théorème de la limite centrale:

$$\sqrt{n}\ M_{\theta_0}(X_1, \ldots, X_n) - \mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial \theta} \log f_{\theta_0}(X_i)\right] \xrightarrow[n \to +\infty]{\mathcal{L}} N_d(0, I_1(\theta_0)),$$

d'après la première définition de l'information de Fisher comme $\mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial \theta} \log f_{\theta_0}(X_i)\right] = 0$ (voir les propriétés précédentes), on obtient la normalité asymptotique de θ_n. ■

Remarquons sous ces hypothèses, l'estimateur du maximum de vraisemblance est bien asymptotiquement sans biais et efficace. Cependant, dans le cas fini il faut avoir un biais et ne peut pas être un estimateur efficace.

3.6 Régions de confiance

En pratique, estimer un paramètre ne suffit pas. On aimerait pouvoir mesurer quelle marge d'erreur on a sur la connaissance de ce paramètre.

Définition. On se place dans le cadre d'un modèle paramétrique $((\Omega, A, I\!P_\theta, \theta \in \Theta), \theta \subset I\!R^p$. Soit $\alpha \in]0, 1[$ un nombre fixé à priori. On appelle région de confiance du paramètre θ de niveau $1-\alpha$ un sous-ensemble aléatoire $R_{1-\alpha}$ inclus dans $I\!R^p$ et défini sur $((\Omega, A_{I\!N}), \text{tel que pour tout } \theta \in \Theta, \{(x_1, \ldots, x_n) \in (\Omega)^{I\!N}, \theta \in R_{1-\alpha}(x_1, \ldots, x_n)\} \in A$ et:

$$\inf_{\theta \in \Theta} \{I\!P_\theta(\theta \in R_{1-\alpha})\} \geq 1 - \alpha. \tag{7}$$

Si une réalisation observée $(X_1(\omega), \ldots, X_n(\omega))$ est connue, $R_{1-\alpha}(X_1(\omega), \ldots, X_n(\omega))$ est appelée région de confiance observée. Dans le cas où le paramètre est une réel ($p = 1$), on pourra obtenir un intervalle de confiance.

Comment déterminer une région de confiance ? En premier lieu, il est clair que $\pi(\theta_n, \theta_0) \in [0, +\infty[$, $\mathbb{R}_+$ ou $\mathbb{R}$ (en général, on choisit π proche de 0, et en particulier $\pi(\theta, \theta) = 0$). Une méthode possible pour la construction de région de confiance est la suivante : naturellement, on utilise un estimateur T convergent de θ, mais souvent également de θ ce qui rend difficile (à quelques exceptions) son utilisation directe. On préférera donc utiliser ce que l'on appelle une fonction pivotale $\pi(T, \theta)$, une fonction mesurable d'un estimateur et de θ dont la statistique limite est connue. On essayera alors de résoudre la propriété (7) sous la forme

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta(\pi(T, \theta) \in C_\alpha) \geq 1 - \alpha,$$

où C_α est une région déterministe. Nous supposerons ensuite construire une région de confiance en fonction des quantiles (souvent à $\alpha/2$ et $1-\alpha/2$) de la loi de la fonction pivotale.

Exemple. Si le modèle est régulier, sous les conditions entraînant la normalité asymptotique du maximum de vraisemblance, on peut également montrer (théorème de Slutski) que

$$\pi(\theta_n, \theta_0) = \sqrt{n} \cdot (I_1(\theta_n))^{1/2} \cdot (\theta_n - \theta_0) \xrightarrow[n \to +\infty]{\mathcal{L}} N_d(0, I_p),$$

où u_d est la matrice identité. En détaillant le produit $(I_1(\theta))^{1/2} \cdot (I_1(\theta))^{1/2} = I_1(\theta)$ pour tout $\theta \in \Theta$. Ainsi, si n est grand, on pourra assimiler la loi de $\pi(\theta_n, \theta_0)$ avec la loi normale centrée réduite multidimensionnelle. Or si $Z \sim N_p(0, I_p)$, avec $q_{1-\alpha/2} > 0$ le quantile d'une normale centrée réduite de niveau $1-\alpha/2$, tel que $P(Z \in [-q_{1-\alpha/2}, q_{1-\alpha/2}]^d) \geq 1 - \alpha$. Ainsi le pavé $\theta_n^{-1/2} \cdot (I_1(\theta_n))^{-1/2} \cdot [-q_{1-\alpha/2}, q_{1-\alpha/2}]^d$ recentré autour de θ_n forme alors la région de confiance cherchée.

3.7 M-estimateur

Il s'agit ici de généraliser les estimateurs du maximum de vraisemblance. On donnera ainsi des critères généraux pour la consistance et la normalité asymptotique des estimateurs.

3.7.1 Introduction

Supposons que nous voulons estimer θ relié à la loi de probabilité d'observations $(X_1, \cdots, X_n)$. La méthode pour trouver un tel estimateur est de minimiser une fonction de la forme

$$\theta \to M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i)$$

Les $m_\theta : \mathbb{R}^p \to \mathbb{R}$ sont des fonctions connues.

Exemple. Estimateur du maximum de vraisemblance. Admettons que $(X_1, \cdots, X_n)$ sont i.i.d. de leur loi à p pour densité f_θ. L'estimateur du maximum de vraisemblance $\hat{\theta}_n$ minimise :

$$\theta \to M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} -\log f_\theta(X_i)$$

Exemple : estimateur des moindres carrés. Admettons que $(X_1, Y_1), \cdots, (X_n, Y_n))$ sont i.i.d., $X \in \mathbb{R}^p, Y_i \in \mathbb{R}$ et vérifient l'équation de régression linéaire :

$$Y = \theta^T X + \varepsilon$$

où ε est une variable d'erreur centrée, indépendante de X et de variance finie. L'estimateur des moindres carrés $\hat{\theta}_n$ de θ est alors celui qui minimise :

$$\theta \to M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta^T X_i)^2$$

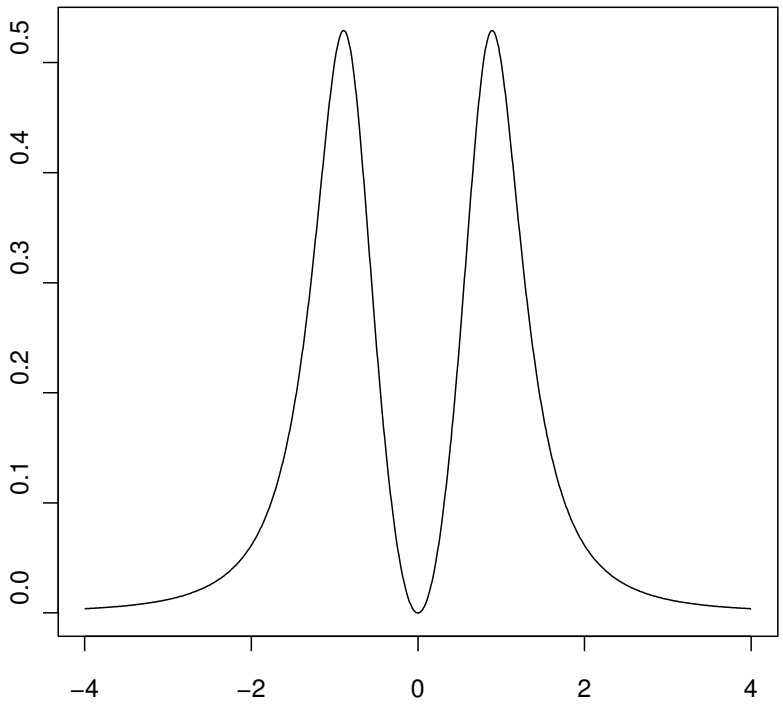

Figure 1 – Minimum mal-séparé

3.7.2 Consistance

Il est important que l'estimateur converge vers la vraie valeur du paramètre à mesure que la taille de l'échantillon augmente indéfiniment ou bien en probabilité, lorsque le nombre d'observations n converge vers l'infini. Si c'est le cas, l'estimateur est dit asymptotiquement consistant. Par exemple la moyenne $\bar{X}$ est asymptotiquement consistante pour la moyenne de la population $\theta = E(X)$ si $E(X)$ existe. On veut donc prouver que:

$$\hat{\theta}_n \xrightarrow{p.s.} \theta$$

On suppose que le M-estimateur $\hat{\theta}_n$ minimise la fonction $M_n(\theta)$. Clairement, le comportement asymptotique de $\hat{\theta}_n$ dépend du comportement asymptotique de la fonction $M_n(\theta)$. Pour une bonne normalisation il existera une fonction $M(\theta)$ déterministe telle que:

$$\forall \theta : M_n(\theta) \xrightarrow{p.s.} M(\theta)$$

Il semblerait raisonnable que le minimisateur $\hat{\theta}_n$ de $M_n(\theta)$ converge sous des conditions raisonnables vers θ_0 le minimisateur de $M(\theta)$. Il faut quand même faire attention à ce que le comportement asymptotique de la minimisation de $M(\theta)$ soit bien "proche" de celui de $M_n(\theta)$. La figure 1 donne un exemple d'une fonction qui a bien un minimum égal à 0, mais la droite d'équation $y = 0$ est aussi une asymptote.

Théorème Soit Θ l'espace des paramètres possibles et M_n, M des fonctions telles que pour tout $\varepsilon > 0$:

$$\begin{aligned} &\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p.s.} 0, \\ &\inf_{\theta : \ \|\theta - \theta_0\| \geq \varepsilon} M(\theta) > M(\theta_0) \end{aligned}$$

alors pour toute suite d'estimateurs $\hat{\theta}_n$ telle que

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} M_n(\theta)$$

On aura

$$\hat{\theta}_n \xrightarrow{p.s.} \theta_0$$

Démonstration: On a pour tout voisinage V de θ_0 l'existence d'une constante $\eta(V)$ telle que:

$$\forall \theta \in \Theta \backslash V,\ M(\theta) > M(\theta_0) + \eta(V)$$

Donc, pour montrer la consistance forte, il suffit de montrer que pour tout V de θ_0 un voisinage

$$\lim_{n\to\infty} \hat{\theta}_n \overset{p.s.}{\subset} V \Leftrightarrow \lim_{n\to\infty} M(\hat{\theta}_n) - M(\theta_0) \overset{p.s.}{\leq} \eta(V)$$

Par définition $M_n(\hat{\theta}_n) \overset{p.s.}{\leq} M_n(\theta_0)$ et comme $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p.s.} 0$ on aura

$$\lim_{n\to\infty} M_n(\hat{\theta}_n) \overset{p.s.}{\leq} M(\theta_0) + \frac{\eta(V)}{2}.$$

De même on aura

$$\lim_{n\to\infty} M(\hat{\theta}_n) - M_n(\hat{\theta}_n) \overset{p.s.}{=} 0$$

et

$$\lim_{n\to\infty} M(\hat{\theta}_n) - \frac{\eta(V)}{2} \overset{p.s.}{<} \lim_{n\to\infty} M_n(\hat{\theta}_n) \overset{p.s.}{<} M(\theta_0) + \frac{\eta(V)}{2}$$

finalement $\lim_{n\to\infty} M(\hat{\theta}_n) - M(\theta_0) \overset{p.s.}{\leq} \eta(V)$ ce qui prouve la consistance forte du M-estimateur.

Conditions suffisantes pour le théorème Pour un modèle régulier les hypothèses du théorème sont facilement vérifiés. Pour obtenir la condition:

$$\inf_{\theta:\ \theta-\theta_0\ \geq\varepsilon} M(\theta) > M(\theta_0)$$

Il suffit que la fonction limite $M(\theta)$ soient une fonction strictement concave et que pour tout $\theta = \theta_0$, $M(\theta) = M(\theta_0)$. Ce la se rapelle si $M(\theta)$ est l'espérance de l'opposé la log-vraisemblance, c'est-à-dire la distance de Kullback à une constante près:

$$M(\theta) = \int f_{\theta_0} \log \frac{f_{\theta_0}}{f_\theta}$$

Pour obtenir l'hypothèse de la loi uniforme des grands nombres:

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p.s.} 0$$

une condition suffisante est que l'ensemble des paramètres possibles Θ soit compact, que la fonction $m_\theta(x)$ soit continue pour tout x et qu'il existe une fonction intégrable $d(x)$ telle que $|m_\theta(x)| \leq d(x)$ pour tout $\theta \in \Theta$.

Preuve Pour une boule ouverte B de Θ, on pose $m^B = \sup_{\theta\in B} m_\theta$ et $m_B = \inf_{\theta\in B} m_\theta$. Par le théorème de convergence dominée $E[m^B - m_B] \to 0$ lorsque le diamètre de la boule tend vers 0. Pour $\varepsilon > 0$, soit $B^1, \cdots, B^k$ un recouvrement fini de Θ tel que $E[m^{B^i}(X)] - m_{B^i}(X)] < \varepsilon$. Pour tout $\theta \in B^i$, on aura:

$$\begin{aligned} M_n(\theta) - M(\theta) &\leq \frac{1}{n} m^{B^i}(X_i) - E[m_{B^i}(X)] \leq \frac{1}{n} m^{B^i}(X_i) - E[m^{B^i}(X)] + \varepsilon \\ M_n(\theta) - M(\theta) &\geq \frac{1}{n} m_{B^i}(X_i) - E[m^{B^i}(X)] \geq \frac{1}{n} m_{B^i}(X_i) - E[m_{B^i}(X)] - \varepsilon \end{aligned}$$

Ainsi,

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \leq \sup_{i\in\{1,\cdots,n\}} \max\left(\frac{1}{n} m^{B^i}(X_i) - E[m^{B^i}(X)],\ \frac{1}{n} m_{B^i}(X_i) - E[m_{B^i}(X)]\right) + \varepsilon$$

Ainsi, presque-sûrement $\forall \varepsilon > 0$, $\lim_{n\to\infty} \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| < \varepsilon$ ■

3.7.3 Normalité Asymptotique

On donne ici des conditions suffisantes pour la normalité asymptotique du M-estimateur. Ces conditions peuvent être affaiblies mais celles données ci-dessous conviennent pour les exemples étudiés précédemment.

Notations

– Pour une suite de variables aléatoires X_n et une suite de réels R_n:

$$X_n = o_P(R_n) \Leftrightarrow X_n = Y_n R_n \text{ et } Y_n \xrightarrow{P} 0$$

– Pour une suite de variables aléatoires X_n et une suite de réels R_n:

$$X_n = O_P(R_n) \Leftrightarrow X_n = Y_n R_n \text{ et } Y_n \text{ est bornée en probabilité à l'ordre près, i.e. } \forall \varepsilon > 0, \exists M \text{ tel que } P(|Y_n| > M) < \varepsilon \text{ pour } n \text{ assez grand.}$$

Hypothèses On suppose que les observations $(X_1, \ldots, X_n)$ sont i.i.d. et que les hypothèses suivantes sont vérifiées:

H1 Le M-estimateur $\hat{\theta}_n$ est fortement consistant

H2 Il existe un voisinage V du vrai paramètre θ_0 tel que pour tout $\theta \in V$, la dérivée seconde par rapport à θ de $m_\theta(X)$ est dominée par une fonction g intégrable.

H3 Le carré de la dérivée 1ère: $\left(\frac{\partial m_\theta(X)}{\partial \theta}(\theta_0)\right)^2$ est intégrable.

H4 La matrice des dérivées secondes $\frac{\partial^2 m_\theta(X)}{\partial \theta^2}(\theta_0)$ est intégrable et inversible.

On aura alors le théorème suivant:

Théorème Sous les hypothèses H 1,...,H 4:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left[-E\left[\frac{\partial^2 m_\theta}{\partial \theta^2}(\theta_0)\right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial m_\theta}{\partial \theta}(\theta_0)\right] + o_P(1)$$

En particulier la suite $\sqrt{n}(\hat{\theta}_n - \theta_0)$ est asymptotiquement normale de moyenne 0 et de matrice de variance-covariance $E\left[\frac{\partial^2 m_\theta}{\partial \theta^2}(\theta_0)\right]^{-1} E\left[\frac{\partial m_\theta}{\partial \theta}(\theta_0)\frac{\partial m_\theta}{\partial \theta}(\theta_0)^T\right] E\left[\frac{\partial^2 m_\theta}{\partial \theta^2}(\theta_0)\right]^{-1}$

Démonstration: Par un développement de Taylor il existe un vecteur aléatoire $\tilde{\theta}_n$ situé sur le segment $[\hat{\theta}_n, \theta_0]$ tel que:

$$0 = \frac{\partial M_n(\hat{\theta}_n)}{\partial \theta} = \frac{\partial M_n(\theta_0)}{\partial \theta} + \frac{\partial^2 M_n(\theta_0)}{\partial \theta^2}(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^T \frac{\partial^3 M_n(\tilde{\theta}_0)}{\partial \theta^3}(\hat{\theta}_n - \theta_0)$$

Le premier terme à droite est la moyenne d'une variable aléatoire $\frac{\partial M_n(\theta_0)}{\partial \theta}$ qui a pour espérance:

$$E\left[\frac{\partial M_n(\theta_0)}{\partial \theta}\right] = 0$$

Par le théorème de la limite centrale la suite $\sqrt{n}\frac{\partial M_n(\theta_0)}{\partial \theta}$ converge en loi vers une gaussienne de moyenne 0 et de matrice de variance-covariance $E\left[\frac{\partial m(\theta_0)}{\partial \theta}\frac{\partial m(\theta_0)}{\partial \theta}^T\right]$. Par la loi des grands nombres $\frac{\partial^2 M_n(\theta_0)}{\partial \theta^2}$ converge presque-sûrement vers une matrice $E\left[\frac{\partial^2 m(\theta_0)}{\partial \theta^2}\right]$. Si k est la dimension du vecteur θ, la dérivée $\frac{\partial^3 M_n(\theta_0)}{\partial \theta^3}$ est un vecteur de k matrices $k \times k$, par hypothèse il existe un voisinage V de θ_0 tel que $\frac{\partial^3 M_n(\theta_0)}{\partial \theta^3}$ est dominée par une fonction g intégrable. Comme $\hat{\theta}_n$ est consistant, $\lim \hat{\theta}_n \in V$ presque-sûrement et pour $\hat{\theta}_n \in V$ on aura le développement suivant:

$$-\frac{\partial M_n(\theta_0)}{\partial \theta} = \left[J + o_P(1)\right] + \frac{1}{2}(\hat{\theta}_n - \theta_0)O_P(1)\left[(\hat{\theta}_n - \theta_0)\right] = \left(J + o_P(1)\right)(\hat{\theta}_n - \theta_0))$$

car $(\hat{\theta}_n - \theta_0)O_P(1) = o_P(1)O_P(1)$ si $\hat{\theta}_n$ converge presque-sûrement vers θ_0. Par ailleurs la probabilité que la matrice $J + o_P(1)$ soit inversible tend vers 1. En multipliant l'égalité précédente par $\sqrt{n}(J + o_P(1))^{-1}$ on obtient le résultat annoncé ■

4 Tests paramétriques

4.1 Principes d'un test

Un test permet, à partir d'une réalisation d'un échantillon, de décider entre deux hypothèses, en mettant en avant une hypothèse privilégiée, appelée hypothèse nulle, et une hypothèse alternative, appelée H_1. On associe à un test un niveau α (avec souvent $\alpha = 0.05$) et une puissance $1-\beta$. La plupart du temps, α est fixé a priori et β s'en déduit. Plus précisément,

Définition. On se place dans le cadre d'un problème paramétrique dominé $(\Omega, \mathcal{A}, I\!P_\theta, \theta \in \Theta)$, où $\Theta \subset I\!R^p$ et soit θ la "vraie" valeur du paramètre. Un problème de test est un choix entre deux hypothèses

$$\begin{cases} H_0 : \theta \in \Theta_0 & : \text{ hypothèse dite nulle} \\ H_1 : \theta \in \Theta_1 & : \text{ hypothèse dite alternative,} \end{cases} \tag{8}$$

où $\Theta_0 \subset I\!R^p$, $\Theta_1 \subset I\!R^d$ et $\Theta_0 \cap \Theta_1 = \emptyset$.

Ce choix, on peut le faire parmi différents types de problèmes de tests suivant les constitutions respectives de Θ_0 et Θ_1.

Définition. Une hypothèse (H_0 ou H_1) est dite simple si elle est associée à un singleton inclus dans (Θ_0 ou Θ_1). Sinon, elle sera dite composite. Dans le cadre ($\Theta \subset I\!R$), si H_0 est simple de la forme $\theta = \theta_0$ et si H_1 est composite de la forme $\theta > \theta_0$ ou $\theta < \theta_0$, on parlera de test unilatéral, si H_1 est composite de la forme $\theta \neq \theta_0$, on parlera de test bilatéral.

Comment faire pour choisir entre les deux hypothèses H_1 et H_2 ? Il faudra partir de ce que l'on peut connaître du modèle, c'est-à-dire généralement un échantillon observé $(X_1, \ldots, X_n)$. Pour cela, on définit une statistique qui servira de clé de voûte du test:

Définition. Dans le cadre du problème de test (8, soit T une statistique (donc une fonction mesurable d'un échantillon $(X_1, \cdots, X_n)$ is sudu modèle) à valeurs dans $I\!R^d$ qui sera appelée statistique du test. Le test sera défini par la fonction $\varphi = I\!I_{T \in W}$, où W est une partie de $I\!R^d$ appelée région critique du test (et partie complémentaire dans $I\!R^d$ sera appelée région d'acceptation du test). Si $\varphi = 1$, on choisira donc on décidera plutôt H_0.

Donc, à chaque hypothèse H_0 et H_1, on associe une partie de $I\!R^d$ pour la statistique de test T et, en général, ces parties ne sont pas Θ_0 et Θ_1. Pour pouvoir déterminer la région W, dans un cadre théorique (qui n'est pas le même que le cadre pratique, voir plus bas), on peut commencer par associer une fonction puissance à la statistique de test, puis les différentes erreurs et enfin une espérance β:

Définition. Pour la statistique de test T, on associe:

- une fonction puissance, qui est la probabilité d'accepter H_1: $\theta \in \Theta_1 \to I\!P_\theta(T \in W)$.
- une erreur de première espèce : $P_{H_0}(\text{Choisir } H_1) = \alpha = \sup_{\theta \in \Theta_0} I\!P_\theta(T \in W)$;
- une erreur de seconde espèce : $P_{H_1}(\text{Choisir } H_0) = \beta = \sup_{\theta \in \Theta_1} I\!P_\theta(T \notin W)$.

La puissance du test est $1-\beta$.

Cependant, ce qui vient d'être écrit relève de la théorie. En pratique, on utilise plutôt la démarche suivante:

Construction d'un test: On suppose le problème de test (8). On pose généralement a priori α qui dépend du problème posé, mais est généralement $\alpha = 0.05$), $1-\alpha$ est appelé niveau du test. Par la suite, on réalise:

1. L'expression quantitative des hypothèses H_0 et H_1.
2. Le choix de la statistique T du test.
3. La construction d'une région critique W à l'hypothèse H_0 par rapport à T.
4. La détermination explicite de W en fonction de α.
5. Le calcul (si possible) de la puissance $1-\beta$ du test.
6. Pour la réalisation de l'échantillon, rejet ou acceptation de H_0.

Remarque : Cependant, en pratique on ne procède pas ainsi. On a donc deux types d'erreur. Le choix de l'hypothèse privilégiée est donc fonction de la gravité relative de ces ultats d'un test n'est pas symétrique. Par exemple, supposons que l'on ait pour modèle $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathcal{N}(\theta, 1)^{\otimes n}, \theta \in \mathbb{R})$ et que l'on veuille tester $H_0 : \theta = 0$ contre $H_1 : \theta = 1$. À partir d'un échantillon $(X_1, \cdots, X_n)$ du modèle, nous verrons plus loin que $\overline{X}_n$ est une statistique de test pertinente. Par exemple, si $\alpha = 0.05$, $\overline{X}_n(\omega) = 1$, et $X_1(\omega) = 0.8$, que va-t-on choisir entre H_0 et H_1 ? Naturellement, une région critique sera de la forme $[s, +\infty[$, où $s \in \mathbb{R}$, car $\overline{X}_n$ est un estimateur de θ. On détermine s à l'aide de la quantité $P_0(\text{Choisir } H_1) = \alpha = P_0(X_1 \geq s)$, donc par exemple si $\alpha = 0.05$, $s \simeq 1.65$. Par suite, si $X_1(\omega) = 0.8$, on accepte H_0 et l'erreur de seconde espèce est $P_1(X_1 < s) \simeq 0.74$, donc nettement élevé : le test n'est pas discriminant. Mais intuitivement, si on inverse H_0 et H_1, c'est-à-dire $H_0 : \theta = 1$ contre $H_1 : \theta = 0$, le même résultat $X_1(\omega) = 0.8$, conduit à accepter avec une erreur de seconde espèce égale à 0.74. On obtient donc deux résultats opposés pour la même expérience aléatoire. Les hypothèses H_0 et H_1 ne sont clairement pas interchangeables.

La question qui se pose maintenant est de savoir comment trouver une statistique de test. Intuitivement elle dépendra du problème testé.

4.2 Test de Wald

Un estimateur du maximum de vraisemblance permet d'associer à chaque échantillon une valeur du paramètre dans l'ensemble de même "forme" que Θ_1. Cependant, la difficulté est de trouver la loi de l'estimateur du maximum de vraisemblance θ à n fixé. Si cela est possible, on utilisera directement θ comme statistique de test.

Sinon, de manière plus générale, on connaît la loi asymptotique de θ quand le modèle est régulier. Donc quand n est grand, on pourra utiliser une loi normale comme approximation. Mais, une autre valeur de θ obstacle apparaît : la matrice de covariance asymptotique, qui est la matrice d'information de Fisher inverse, dépend du paramètre θ. Aussi va-t-on préférer utiliser la statistique de test T suivante :

Définition. Pour un modèle paramétrique dominé régulier $(\Omega, \mathcal{A}_n, I\!P_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$. La statistique de Wald T pour le test $H_0 : \theta = \theta_0$ contre $H_1 : \theta \in \Theta_1$ est : $T_n = n \cdot {}^t(\theta_n - \theta) \cdot I(\theta) \cdot (\theta_n - \theta)$.

Pour montrer "heuristiquement" la pertinence de ce test, nous considérons la suite de tests (T_n) se plaçant dans le "grand" modèle asymptotique :

Théorème. Dans le cadre d'un modèle paramétrique $((\Omega^{\mathbb{N}}, \mathcal{A}_{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, domine par une mesure σ-finie, pour le problème de test $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_0$, alors, en notant T_n la statistique de Wald pour le modèle de taille n sous l'hypothèse H_0,

$$T_n \xrightarrow[n \to +\infty]{L} \chi^2(p).$$

La région de rejet asymptotique du test sera donc de la forme $T_n > \chi^2_\alpha$, où χ^2_α est le quantile d'ordre $1-\alpha$ de la loi du $\chi^2(p)$. La suite de tests (T_n) a donc une puissance qui tend vers 1 lorsque n est fixé.

Démonstration : La loi asymptotique de θ induit la loi asymptotique de T, car $\sqrt{n} \cdot I(\theta)^{1/2} \cdot (\theta_n - \theta)$ suit asymptotiquement une $\mathcal{N}(0, I_d)$ sous l'hypothèse H_0 et $T_n = \sqrt{n} \cdot I(\theta)^{1/2} \cdot (\theta_n - \theta)^2$. ■

Voici donc un premier type de test qui, sous certaines conditions de régularité du modèle et pour certaines hypothèses de tests existant. Mais pour rait-on faire mieux ? Et en quel sens ? Nous faut donc définir un moyen de comparaison entre deux tests.

4.3 Test du rapport de vraisemblance

Définition. Sous les hypothèses et notations précédentes, on dira qu'un test φ est uniformément le plus puissant (U.P.P.) au seuil α si le niveau de φ associé à la statistique T est inférieur ou égal à α et si pour tout autre test φ' associé à la statistique T' de niveau inférieur ou égal à α, $\forall \theta \in \Theta_1$,

$$I\!E_\theta(\varphi) = 1 - I\!P_\theta(T \in W) \leq 1 - I\!P_\theta(T' \in W') = I\!E_\theta(\varphi').$$

Définitions Sous les hypothèses précédentes, si L_θ est la vraisemblance, on appellera test du rapport de vraisemblance (test de Neyman - Pearson dans le cas d'hypothèses simples) un test de statistique T telle que

$$T = \frac{\sup_{\theta \in \Theta_0} L_\theta(X_1, \cdots, X_n)}{\sup_{\theta \in \Theta_1} L_\theta(X_1, \cdots, X_n)}.$$

La région critique W associée à un tel test est de la forme $W =]-\infty, K[$ (donc si $T < K$, on rejette H_0).

Une des vertus du test du rapport de vraisemblance par rapport au test de Wald est qu'il peut être utilisé dans un modèle non régulier (mais la question de sa loi, ou de la loi d'une fonctionnelle de ce test, demeure entière). De plus, la propriété suivante confirme l'intérêt de cette statistique de test:

Propriété (Principe de Lehmann) Dans le cas du test de deux hypothèses simples, ou d'un test unilatéral ($\Theta \subset \mathbb{R}$), ce test est U.P.P. Dans le cas d'un test bilatéral, il n'existe pas forcément de test U.P.P.

Démonstration : ■

Enfin, un tel test pour un modèle régulier, va pouvoir être mis en œuvre grâce à la normalité asymptotique de l'estimateur du maximum de vraisemblance:

Théorème Dans le cadre d'un modèle paramétrique $((\Omega, A)^{\otimes \mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, domination par une mesure régulière, pour le problème de test $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_0$, alors, en notant T_n la statistique du rapport de vraisemblance pour un échantillon de taille n,

$$-2 \log(T_n) \xrightarrow[n \to +\infty]{L} \chi^2(p).$$

La région de rejet asymptotique du test sera donc $-2 \log(T_n) > s_\alpha$, où s_α est le quantile d'ordre $1 - \alpha$ de la loi du $\chi^2(p)$. La suite de tests (T_n) a donc une puissance qui tend vers 1 lorsque α est fixé.

Démonstration : la démonstration reprend un peu celle de la normalité asymptotique du maximum de vraisemblance. ■

Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2005-2006

Statistique

Contrôle continu n°1, novembre 2005

Examen de 2 h 00. Tout document ou calculatrice est interdit.

1. On considère une suite $(X_i)_{i\in\mathbb{N}}$ de variables aléatoires indépendantes et identiquement distribuées suivant une $\mathcal{N}(m,\sigma^2)$, où $m\in\mathbb{R}$ et $\sigma^2>0$ sont des paramètres inconnus.
   Soit également pour $n\in\mathbb{N}^*$, $\overline{X}_n=\frac{1}{n}(X_1+\cdots+X_n)$ et $\sigma_n^2=\frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X}_n)^2$.

   (a) Pour n fixé, déterminer le modèle statistique paramétrique.

   (b) Pour $n\in\mathbb{N}^*$, quelle est la loi de $\overline{X}_n$?

   (c) Quelles sont les limites (en probabilité ou en loi) de $\overline{X}_n$ et de σ_n^2 (justifier ...)?

   (d) Montrer que la connaissance de $\overline{X}_n$ et de $(\overline{X}_n)$ induit celle de $(X_1,\cdots,X_n)$. Déterminer la loi du vecteur $(X_1,\cdots,X_n)$. Les (X_k) sont-elles indépendantes?

   (e) Soit $\overline{\overline{X}}_n=\frac{1}{n}(\overline{X}_1+\cdots+\overline{X}_n)$. Quelle est la loi de $\overline{\overline{X}}_n$ pour $n\in\mathbb{N}^*$? En déduire que $\overline{\overline{X}}_n \xrightarrow[n\to+\infty]{P} m$. Montrer également que $\overline{\overline{X}}_n \xrightarrow[n\to+\infty]{p.s.} m$.

   (f) Comment peut-on faire pour savoir quelle suite de variables aléatoires $(\overline{X}_k)_{k\geq 1}$ ou $(\overline{\overline{X}}_k)_{k\geq 1}$, s'approche le plus vite de m ? Conclusion?

   (g) Pour le modèle paramétrique de taille n est supposé connue, montrer que la statistique $\overline{X}_n$ est exhaustive comme et la statistique $(X_1,\cdots,X_n)$? Enfin, la statistique $\overline{\overline{X}}_n$ est-elle exhaustive?

2. En fait, on ne connaît pas explicitement chaque variable mais plutôt pour tout $k\in\mathbb{N}^*$, $T_k=\max(X_1,\cdots,X_k)$.

   (a) La connaissance de $(T_1,\cdots,T_n)$ induit-elle celle de $(X_1,\cdots,X_n)$?

   (b) Déterminer la fonction de répartition de T_k, puis, après avoir montré son existence, sa densité par rapport à la mesure de Lebesgue, le tout en fonction de la fonction de répartition F et de la densité de X_1.

   (c) Déterminer, en justifiant, le comportement asymptotique (quand $n\to+\infty$) de $(T_n)_n$.

   (d) Pour $k\in\mathbb{N}^*$, montrer que T_k et T_{k+1} ne sont pas indépendantes. Montrer que $\mathbb{P}(T_{k+1}=T_k)=\frac{k}{k+1}$. En déduire la mesure de probabilité de la variable $T_{k+1}-T_k$. La loi de probabilité de la variable T_k est-elle "continue" ? "discrète" ?

   (e) La statistique T_n est-elle exhaustive pour le modèle paramétrique de taille n où σ^2 est supposée connue ? Et la statistique $(T_1,\cdots,T_n)$?

Université Pierre et Marie Curie - Sorbonne

Première Année Master M.A.E.F. 2005 – 2006

Statistique

Contrôle continu n°2, janvier 2006

Examen de 2h00. Tout document ou calculatrice est interdit.

1. Soit la variable X qui suit une loi dont la densité par rapport à la mesure de Lebesgue sur]0, 1] est, avec $\beta \in \mathbb{R}$:

$$f_X(x) = K \cdot x^{\beta} \quad \text{pour tout } x \in]0, 1],$$

   (a) Déterminer K en fonction de β, en précisant quelle condition on doit imposer à β. En déduire $\mathbb{E}(X)$ et $\mathrm{var}(X)$, en précisant également les conditions sur β.

   (b) On suppose que la suite $(X_i)_{i \in \mathbb{N}^*}$ est constituée de variables aléatoires indépendantes et identiquement distribuées suivant la même loi que X. Soit un échantillon observé $(X_1, \ldots, X_n)$. On désire estimer β à partir de cet échantillon. Quel est le modèle statistique ? Montrer que ce modèle appartient à la famille exponentielle.

   (c) En déduire qu'il n'existe pas d'estimateur sans biais efficace de β.

   (d) Montrer que $-\log(X)$ suit une loi connue donnée par le paramètre $\frac{1}{\beta+1}$. En déduire que $\tilde{\beta}_n = -1 - n \cdot \left(\sum_{i=1}^{n} \log(X_i) \right)^{-1}$ est un estimateur sans biais de β (utiliser les lois gammas...), puis qu'il est de variance uniformément minimale parmi les estimateurs sans biais (Lehmann-Scheffé).

2. Soit Y une variable suivant une loi de Bernoulli de paramètre $p \in]0, 1[$ et indépendante de X. On définit une variable Z de la manière suivante: si $Y = 1$, alors $Z = X$, et si $Y = 0$ alors $Z = -X$.

   (a) Montrer que Z suit une loi absolument continue par rapport à la mesure de Lebesgue sur $[-1, 0[\cup]0, 1]$ et que sa densité est:

$$f_Z(z) = (\beta + 1)|x|^{\beta} \, p \cdot \mathbb{1}_{x \in]0,1]} + (1-p) \cdot \mathbb{1}_{x \in [-1, 0[} \quad \text{pour tout } x \in [-1, 0[\cup]0, 1].$$

   Calculer $\mathbb{E}(Z)$ et $\mathrm{var}(Z)$ (en précisant les conditions sur β).

   (b) On suppose que la suite $(Z_i)_{i \in \mathbb{N}^*}$ est constituée de variables aléatoires indépendantes et identiquement distribuées suivant la même loi que Z. Soit un échantillon observé $(Z_1, \ldots, Z_n)$. On désire estimer (β, p) à partir de cet échantillon. Quel est le modèle statistique ? Montrer que ce modèle appartient à la famille exponentielle.

   (c) En déduire une statistique exhaustive minimale pour ce modèle. Déterminer la matrice d'information de Fisher de ce modèle, puis la borne de Cramer-Rao. Déterminer une fonction g de (β, p) que l'on peut estimer sans biais et de manière efficace.

(d) Déterminer après avoir montré l'unicité, l'estimateur (β_n, p_n) du maximum de vraisemblance de (β, p). Les estimateurs β_n et p_n sont-ils indépendants ? Déterminer un théorème de la limite centrale pour (β_n, p_n). Est-ce un estimateur asymptotiquement efficace ?

(e) Déterminer une région de confiance de niveau 95 % sur (β, p), en utilisant 1/ l'estimateur efficace de $g(\beta, p)$ 2/ l'estimateur de maximum de vraisemblance dans le cadre asymptotique.

Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2005 - 2006

Statistique

Examen terminal, janvier 2006

Examen de 3 h 00. Tout document ou calculatrice est interdit.

1. On considère $(X_k)_{k\in\mathbb{N}}$ et $(X'_k)_{k\in\mathbb{N}}$ deux suites indépendantes de variables aléatoires définies sur le même espace de probabilité, indépendantes et identiquement distribuées suivant les lois respectives $N(\mu,\sigma^2)$ (pour les X_k) et $N(\mu',\sigma^2)$ (pour les X'_k), où $(\mu,\mu')\in\mathbb{R}^2$ et $\sigma^2>0$. Le but du problème est de tester $\mu=\mu'$ à partir d'échantillons de chacune de ces suites.

   Soit $(X_1,\ldots,X_n)$ et $(X'_1,\ldots,X'_{n'})$, où $n\in\mathbb{N}^*$ et $n'\in\mathbb{N}^*$, deux échantillons issus de $(X_k)_{k\in\mathbb{N}}$ et $(X'_k)_{k\in\mathbb{N}}$. On pose $Z=(Z_1,\ldots,Z_{n+n'})=(X_1,\ldots,X_n,X'_1,\ldots,X'_{n'})$

   (a) Déterminer le modèle statistique associé à $Z=(Z_1,\ldots,Z_{n+n'})$.

   (b) Montrer que ce modèle est exponentiel. En déduire que μ et μ' peuvent être estimés efficacement (on notera $\overline{X}_n$ et $\overline{X}'_{n'}$ les estimateurs respectifs).

   (c) Montrer qu'un estimateur du maximum de vraisemblance de σ^2, lorsque $\mu=\mu'$, est:

$$\sigma^2=\frac{1}{n+n'}\left(\sum_{j=1}^{n}\left(Z_j-\mu\right)^2+\sum_{j=n+1}^{n+n'}\left(Z_j-\mu'\right)^2\right).$$

   Est-ce un estimateur de σ^2? Est-il convergent ? Efficace?

   (d) Lorsque n et n' sont "grands", déduire de ce qui précède, des intervalles de confiance à 95 % pour μ et μ'.

   (e) Soit le problème de test:

$$H_0:\ \mu=\mu' \quad \text{contre} \quad H_1:\ \mu\neq\mu'$$

   (σ^2 restant inconnu). Montrer que la statistique T du rapport de vraisemblance vérifie:

$$T=\left(\frac{\sigma^2}{\overline{\sigma}^2}\right)^{(n+n')/2} \quad \text{avec} \quad \begin{cases}\overline{\sigma}^2 &= \dfrac{1}{n+n'}\displaystyle\sum_{j=1}^{n+n'}\left(Z_j-\overline{Z}_{n+n'}\right)^2\\ \overline{Z}_{n+n'} &= \dfrac{1}{n+n'}\displaystyle\sum_{i=1}^{n+n'}Z_i\end{cases}.$$

   En déduire que la région d'acceptation du test est de la forme $\sigma^2\geq K\cdot\overline{\sigma}^2$, avec K dépendant du niveau du test.

(f) Pour déterminer la valeur K en fonction du niveau $1-\alpha$ du test, on peut considérer la statistique, dite de Fisher,

$$T' = \frac{(n+n')\cdot\overline{\sigma}^2 - (n+n')\cdot\sigma^2}{\frac{n+n'}{n+n'-2}\ \sigma^2}$$

i. Soit les vecteurs de $IR^{n+n'}$, $u_1 = (1,\dots,1)$, $u = (1,\dots,1,0,\dots,0)$ (soit n fois 1 et n' fois 0) et $u' = u_1 - u$. Montrer que u et u' sont orthogonaux.

ii. Sous l'hypothèse H_0, déterminer une expression plus simple de $P_{<u_1>}(Z)$, projeté orthogonal de Z sur le sous-espace vectoriel (s.e.v.) engendré par u_1, et $P_{<u,u'>}(Z)$, projeté orthogonal de Z sur le s.e.v. engendré par u et u'.

iii. Montrer que sous l'hypothèse H_0, le vecteur Z peut s'écrire $Z = \mu u_1 + \sigma\cdot\varepsilon$, où ε est un vecteur aléatoire gaussien composé de $n+n'$ variables gaussiennes centrées réduites.

iv. Montrer que sous l'hypothèse H_0, $(n+n')\cdot\sigma^2 = \sigma^2\cdot\|P_A(\varepsilon)\|^2$, où $\|.\|$ est la norme euclidienne classique sur $IR^{n+n'}$ et A est un s.e.v. de $IR^{n+n'}$ que vous préciserez.

v. En utilisant le théorème de Pythagore, montrer que sous l'hypothèse H_0, avec B est s.e.v. de $IR^{n+n'}$ que vous préciserez, $(n+n')\overline{\sigma}^2-(n+n')\cdot\sigma^2 = \sigma^2\cdot\|P_B(\varepsilon)\|^2$.

vi. En utilisant le Théorème de Cochran, montrer que sous l'hypothèse H_0, T' suit une loi de Fisher à $(1,(n+n'-2))$ degrés de liberté. Lorsque n et n' sont "grands", quelle loi suit approximativement T'?

vii. Pour finir, déterminer K en fonction d'un quantile de la loi de Fisher à $(1,(n+n'-2))$ degrés de liberté.

2. Soit X une variable aléatoire dont la mesure de probabilité est absolument continue par rapport à la mesure de Lebesgue sur IR et de densité:

$$f(x) = \frac{\lambda}{2}\exp(-\lambda|x-m|) \quad \text{pour } x\in IR,$$

avec $m\in IR$ et $\lambda>0$, des paramètres inconnus.

(a) Calculer l'espérance et la variance de X.

(b) Calculer $IP(X=m)$ et $IP(X<m)$. En déduire la médiane (théorique) de la loi de X.

(c) Soit une suite $(X_i)_{i\in IN}$ de variables aléatoires indépendantes et identiquement distribuées suivant la même loi que X, dont on extrait un échantillon observé $(X_1,\dots,X_{2n+1})$. Par ailleurs, on note $X_{(1)}\le X_{(2)}\le\dots\le X_{(2n+1)}$ la statistique d'ordre associée. Soit:

$$H_n(a) = \frac{1}{2n+1}\sum_{i=1}^{2n+1}|X_i-a| \quad \text{pour } a\in IR.$$

Calculer $H_n(X_{(n+1)})$ en fonction des $X_{(i)}$. Montrer que la fonction $a\mapsto H_n(a)$ est minimale en $X_{(n+1)}$ (on pourra développer $H_n(X_{(n+k)})$ en fonction des $X_{(i)}$ pour $k>1$).

(d) On suppose ici que $m = 1$, donc que m est connu ($\lambda > 0$ restant inconnu). Quel est alors le modèle statistique ? Montrer que ce modèle appartient à la famille exponentielle, et en déduire une statistique exhaustive dont vous montrerez qu'elle est complète. Déterminer la matrice d'information de Fisher du modèle. Quelle est la fonction de λ (à une transformation affine près) que l'on peut estimer efficacement? Déterminer l'estimateur de maximum de vraisemblance de λ et montrer qu'il vérifie un théorème de la limite centrale.

(e) On suppose désormais que $m \in \mathbb{R}$ est inconnu, tout comme $\lambda > 0$. Quel est alors le modèle statistique ? Montrer que ce modèle n'appartient pas à la famille exponentielle. A l'aide de la question 2(c), déterminer un estimateur (m_n, λ_n) du maximum de vraisemblance du couple (m, λ).

(f) Pour $a \in \mathbb{R}$, démontrer que $H_n(a)$ converge presque sûrement quand $n \to +\infty$ vers $\mathbb{E}(|X - a|)$. Montrer que la fonction $a \in \mathbb{R} \mapsto \mathbb{E}(|X - a|)$ est minimale en $a = m$. En déduire que $m_n \xrightarrow[n \to +\infty]{p.s.} m$, puisque $\lambda_n \xrightarrow[n \to +\infty]{p.s.} \lambda$.

Université Pierre et Marie Curie - Paris VI

Première Année Master M.A.E.F. 2005 – 2006

Statistique

Examen de septembre 2006

Examen de 3h00. Tout document ou calculatrice est interdit.

1. On considère $(X_k)_{k\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant la loi $\mathcal{N}(m,1)$, où $m\in\mathbb{R}$. Soit la suite de variables $(Y_k)_{k\in\mathbb{N}}$ telle que pour tout $k\in\mathbb{N}^*$,
$$Y_k = X_1 + \cdots + X_k.$$
Soit $(Y_1,\ldots,Y_n)$, où $n\in\mathbb{N}^*$ un échantillon issus de $(Y_k)_{k\in\mathbb{N}}$.
   (a) Déterminer la loi de Y_k pour $k\in\mathbb{N}^*$.
   (b) Déterminer la loi du vecteur $(Y_1,\ldots,Y_n)$. Montrer que pour $i\neq j$, Y_i n'est pas indépendante de Y_j.
   (c) Déterminer le modèle statistique associé à $(Y_1,\ldots,Y_n)$.
   (d) Montrer que ce modèle est exponentiel.
   (e) Soit J_n la matrice de covariance de $(Y_1,\ldots,Y_n)$. Vérifier que
$$J_n^{-1} = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$
   En déduire que m peut être estimé par un estimateur m_n que l'on déterminera sans biais et efficace.
   (f) Déterminer l'estimateur du maximum de vraisemblance de m. Est-ce un estimateur biaisé ? Est-il convergent ? Efficace ? Quel est son risque quadratique?
   (g) Déterminer la statistique du test de rapport de vraisemblance pour le test
$$H_0 : m=m_0 \quad \text{contre } H_1 : m=m_0,$$
où m_0 est une constante connue. On réalise une application numérique de ce test au niveau 5 % pour $n = 1000$. On trouve $m_n = 1.4$. Accepte-t-on alors l'hypothèse H_0 ?

2. Soit X une variable aléatoire dont la mesure de probabilité est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$ et de densité
$$f(x) = k\cdot\frac{1}{\sqrt{x}}\cdot \mathbb{1}_{0\le x\le\theta} \text{ pour } x\in\mathbb{R},$$
avec $k\in\mathbb{R}$ et $\theta > 0$, des paramètres inconnus.

(a) Déterminer l'expression de k en fonction de θ. Calculer l'espérance et la variance de X.

(b) Soit une suite $(X_i)_{i \in \mathbb{N}^*}$ de variables aléatoires indépendantes et identiquement distribuées suivant la même loi que X, dont on extrait un échantillon $(X_1, \ldots, X_n)$. Par ailleurs, on note $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ la statistique d'ordre associée. On désire estimer θ à partir de $(X_1, \ldots, X_n)$. Quel est alors le modèle statistique ? Quel est la vraisemblance du modèle ?

(c) Montrer que $T = X_{(n)} = \max(X_1, \ldots, X_n)$ est une statistique exhaustive pour ce modèle.

(d) Montrer que cette statistique est minimale.

(e) Soit $(x_1, \ldots, x_n) \in \mathbb{R}^n$ et le même n-uplet ordonné $(x) = x_{(1)} \leq \cdots \leq x_{(n)} = \max(x_i)$. Montrer que:

$$\mathbb{P}(X_{(1)} \leq x_{(1)}, \ldots, X_{(n)} \leq x_{(n)}) = n! \mathbb{P}(X_1 \leq x_{(1)} \cap X_1 \leq X_2 \leq x_{(2)} \cap \cdots \cap X_{n-1} \leq X_n \leq x_{(n)}).$$

Montrer par récurrence sur les ordres partielles que:

$$\frac{\partial^n}{\partial x_{(1)} \ldots \partial x_{(n)}} \mathbb{P}(X_1 \leq x_{(1)} \cap X_1 \leq X_2 \leq x_{(2)} \cap \cdots \cap X_{n-1} \leq X_n \leq x_{(n)}) = \prod_{i=1}^{n} f(x_{(i)}).$$

Soit $L_\theta^{(n)}$ la vraisemblance de $(X_1, \ldots, X_{(n)})$. Déduire de ce qui précède que:

$$L_\theta^{(n)}(x_{(1)}, \ldots, x_{(n)}) = n! \cdot L_\theta(x_1, \ldots, x_n).$$

(f) Déterminer la densité puis le biais de T_n.

(g) Montrer que T_n est une statistique exhaustive et complète.

(h) Déduire de ce qui précède un estimateur T_n' de θ, sans biais et uniformément de variance minimale.

(i) Calculer le risque quadratique de T_n' et en déduire que $T_n' \xrightarrow[n \to +\infty]{P} \theta$, puis, que pour tout $\alpha \in [0, 1[$, $n^\alpha (T_n' - \theta) \xrightarrow[n \to +\infty]{P} 0$.

(j) Déterminer explicitement un intervalle de confiance à 95 % de θ en fonction de

Un iv e rs ité Pa ris I , Pa nth éon - So rbo nne

Pr emièr e A nné e M aster M .A .E.F. 20 06 – 20 07

S tati sti qu esI

C o nt rôle cont i nu, 1, novemb r e 2 0 06

Exam en de 2 h 00 . To ut d ocum en t ou ca lcula trice est in terd it.

1. On co nsid ère une suit e $(\varepsilon_i)_{i\in\mathbb{N}}$ de va riables a léatoires ind épendantes et identiquement dist r ibuées suiva nt une loi $N(0,\sigma^2)$, o`u $\sigma^2 > 0$ est un pa ramètre inco nnu. Po ur to ut $n \in \mathbb{N}^*$, on d éfinit:

$$X_n = \varepsilon_n - \alpha\varepsilon_{n-1},$$

o`u $\alpha \in \mathbb{R}$ est inco nnu. On no tera pa r la suite, $\theta = (\sigma^2, \alpha)$.

(a) Mo ntrer que $\mathbb{E}(X_n) = 0$ et $\mathrm{var}(X_n) = (1+\alpha^2)\sigma^2$.

(b) Po ur to ut $n \in \mathbb{N}^*$, d éterminer $\mathbb{E}(X_n X_{n+1})$ et $\mathrm{var}(X_n + X_{n+1})$. Mo ntr er que (X_n) est une suit e de va r ia bles identiquement dist ribu ées dont vo us préc isérez la lo i.

(c) Montrer que $\mathrm{cov}(X_i, X_j) = (1+\alpha^2)\sigma^2$ si $i = j$, $\mathrm{cov}(X_i, X_j) = -\alpha\sigma^2$ si $|i-j| = 1$ et $\mathrm{cov}(X_i, X_j) = 0$ sino n.

(d) Po ur n fix é, en d éduire la loi du vect eur $(X_1, \ldots, X_n)$, puis d éterminer le mod èle statistique pa ramétrique asso cié, précisa nt une mesure do mina nte.

(e) So it $\hat{\sigma}^2_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2$. Mo ntr er que $\hat{\sigma}^2_n$ est un estima teur no n biais é. So it

$$Z_n^{(1)} = \frac{1}{n}\sum_{k=1}^{[n/2]} X_{2k}^2 \quad \text{et} \quad Z_n^{(2)} = \frac{1}{n}\sum_{k=1}^{[(n+1)/2]} X_{2k-1}^2$$

avec $[x]$ la pa rtie ent ière de x. Mo ntrer que les suites de va ria bles $(Z_n^{(1)})_n$ et $(Z_n^{(2)})_n$ co nverg ent presque s ûrement (pour s limit es) et qu'elles v érifient un théorème de la limit e centra le.

(f) Mo ntrer que si deux suites de va ria bles co nvergent presque s ûrement, la suite p osé de leurs so mmes co nverge presque s ûrement. En déduire que $\hat{\sigma}^2_n$ co nverge pr esque s ûrement vers σ^2.

(g) So it $\rho_n = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n-1} X_i X_{i+1}$. En utilisa nt le même typ e d'argument que précédemment, mo ntr er que ρ_n co nverge presque s ûrement vers $-\alpha\sigma^2$. En déduir e un estima teur de α co nverg ea nt presque s ûrement.

2. So it X une varia ble a léatoire dont la loi est absolument continue par rapport à la mesure de Leb esgue sur $\mathbb{R}$ et telle que sa densité par rapp ort à cette mesure soit:

$$f_{\alpha,a}(x) = K \cdot \frac{1}{|x|^\alpha}\mathbb{1}_{-a \le x \le a}$$

avec $a \in \mathbb{R}$ et $\alpha \in \mathbb{R}$. Soit $(X_1, \ldots, X_n)$ un échantillon de n v.a. i.i.d. de même loi que X.

(a) Après avoir précisé l'ensemble des valeurs Θ pour $\theta = (a, \alpha)$, déterminer K en fonction de a et α.

(b) Calculer $\mathbb{E}(X)$ après avoir vérifié que cette intégrale existe effectivement pour $\theta \in \Theta$.

(c) Quel est le modèle statistique associé à $(X_1, \ldots, X_n)$.

(d) Montrer que $S = (|X_1|, \ldots, |X_n|)$ est une statistique exhaustive. Montrer que pour $n \geq 3$ cette statistique n'est pas minimale.

(e) Déterminer une statistique $T = (T_1, T_2)$ à valeurs dans $\mathbb{R} \times [0, +\infty[$ qui soit exhaustive minimale pour tout $\theta \in \Theta$.

Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2006 – 2007

Statistiques I

Contrôle continu n°2, janvier 2007

Examen de 2h00. Tout document ou calculatrice est interdit.

1. Soit X une variable aléatoire suivant la loi suivante:

$$IP(X = 1) = IP(Y = -1) = p \text{ et } IP(X = 0) = 1 - 2p,$$

où p est un paramètre inconnu.

(a) Déterminer l'ensemble des valeurs possibles pour p. Calculer EX et varX.

(b) On suppose que la suite $(X_i)_{i \in IN}$ est constituée de variables aléatoires indépendantes et identiquement distribuées suivant la même loi que X. Soit un échantillon $(X_1, \ldots, X_n)$. Déterminer le modèle statistique associé à cet échantillon et déterminer une mesure dominant ce modèle. Montrer que le modèle appartient à la famille exponentielle. En déduire une statistique exhaustive complète pour ce modèle. Montrer que p peut être estimé efficacement et donner un tel estimateur. Calculer la borne de Cramer-Rao et vérifier qu'elle est bien atteinte par cet estimateur.

(c) On définit la suite $(Y_i)_{i \in IN^*}$ à partir de $(X_i)_{i \in IN}$ de la manière suivante:

$$Y_{i+1} = X_i \cdot X_{i+1} \quad \text{pour } i \in IN.$$

Déterminer la loi de Y_i. Montrer que $cov(Y_i, Y_{i+1}) = 0$. Les $(Y_i)_i$ sont-elles indépendantes?

(d) Montrer que $(|Y_1|, \ldots, |Y_n|)$ est une statistique exhaustive pour le modèle statistique induit par $(Y_1, \cdots, Y_n)$.

2. Soit la variable X qui suit une loi dont la densité f_X par rapport à la mesure de Lebesgue sur IR est, avec $\theta > 0$ et $\alpha > 0$:

$$f_X(x) = K \cdot x^{\alpha} II_{0 \le x \le \theta} \quad \text{pour tout } x \in IR,$$

(a) Déterminer K en fonction de α et θ.

(b) Montrer que $Y = \log(\theta / X)$ suit une loi exponentielle dont on précisera le paramètre.

(c) On suppose que la suite $(X_i)_{i \in IN}$ est constituée de variables aléatoires indépendantes et identiquement distribuées suivant la même loi que X. Soit un échantillon $(X_1, \ldots, X_n)$. On suppose que (θ, α) est inconnu. Préciser alors le modèle statistique formé par cet échantillon et la mesure dominante. Ce modèle appartient-il à la famille exponentielle?

(d) Dans cette question, et uniquement dans cette question, on suppose que θ est connu. Préciser alors le modèle statistique. Ce modèle appartient-il à la famille exponentielle ? Montrer que l'estimateur du maximum de vraisemblance existe, est unique et s'écrit :

$$\tilde{\alpha}_n = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \ln(\theta / X_i)} - 1$$

Montrer que $\tilde{\alpha}_n$ converge presque sûrement vers α et établir un théorème de la limite centrale que l'on exprimera. En déduire un intervalle de confiance à 95% sur α pour n grand.

(e) Dans cette question θ et α sont inconnus. Déterminer une statistique exhaustive pour le modèle. En vous aidant de la question précédente, déterminer l'estimateur du maximum de vraisemblance $(\hat{\theta}_n, \hat{\alpha}_n)$ de (θ, α). Déterminer la fonction de répartition de $\ln(\theta/\theta)$ et en déduire que $\hat{\theta}_n \xrightarrow[n \to +\infty]{P} \theta$, puis que $\sqrt{n} \ln(\theta/\hat{\theta}_n) \xrightarrow[n \to +\infty]{P} 0$.

(f) Soit $(U_n)_{n \in \mathbb{N}}$ et $(V_n)_{n \in \mathbb{N}}$ deux suites de variables réelles définies sur le même espace de probabilité. Montrer que si $(U_n)_n$ converge vers une loi $P(V_n)_n$ converge en probabilité vers 0, alors $(U_n V_n)_n$ converge en loi vers 0 (on pourra par exemple majorer la fonction de répartition adaptée). En déduire que $\tilde{\alpha}_n$ suit le même théorème de la limite centrale que $\tilde{\alpha}_n$

Université Pierre et Marie Curie - Paris VI

Première Année Master M.A.E.F. 2006 - 2007

Statistique

Examen terminal, janvier 2007

Examen de 3 h 00. Tout document ou calculatrice est interdit.

1. On considère une suite de variables aléatoires $(X_k)_{k \in \mathbb{N}^*}$ définies sur le même espace de probabilité, indépendantes et identiquement distribuées suivant une loi de Bernoulli de paramètre p. On définit :

$$Y = \min \{k \in \mathbb{N}^*,\ X_k = 0\}.$$

(a) Comment peut-on interpréter la variable Y ? Montrer que la loi de Y est :

$$\mathbb{P}(Y = k) = p^{k-1}(1-p) \quad \text{pour } k \in \mathbb{N}^*.$$

(b) On suppose que $(Y_1, \ldots, Y_n)$ est un échantillon de variables aléatoires indépendantes et identiquement distribuées suivant la loi de Y avec $p \in]0, 1[$ inconnu. Déterminer le modèle statistique et sa mesure dominante. Montrer que ce modèle est exponentiel. En déduire un estimateur de p sans biais et efficace. Déterminer la borne de Cramer-Rao et vérifier que cette borne est bien atteinte par p.

(c) La variable Y est trop grande lorsqu'elle est trop "grande", par rapport à $T \in \mathbb{N}^*$, c'est-à-dire que l'on définit une variable Z telle que $Z = \min(Y, T)$.

2. Soit X une variable aléatoire réelle dont la mesure de probabilité est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$ et de densité

$$f(x) = \frac{\lambda}{2} \exp(-\lambda |x - m|) \quad \text{pour } x \in \mathbb{R},$$

avec $m \in \mathbb{R}$ et $\lambda > 0$, des paramètres inconnus.

(a) Calculer l'espérance et la variance de X.

(b) Calculer $P(X = m)$ et $P(X < m)$. En déduire une méthode (heuristique) de la loi de X.

(c) Soit une suite $(X_i)_{i \in \mathbb{N}}$ de variables aléatoires indépendantes et identiquement distribuées suivant la même loi que X, dont on extrait un échantillon $(X_1, \ldots, X_{2n+1})$. Par ailleurs, on note $X_{(1)} \le X_{(2)} \le \ldots \le X_{(2n+1)}$ la statistique d'ordre associée. Soit :

$$H_n(a) = \frac{1}{2n+1} \sum_{i=1}^{2n+1} |X_i - a| \quad \text{pour } a \in \mathbb{R}.$$

Calculer $H_n(X_{(n+1)})$ en fonction des $X_{(i)}$. Montrer que la fonction $a \to H_n(a)$ est minimale en $X_{(n+1)}$ (on pourra évaluer $H_n(X_{(n+k)})$ en fonction des $X_{(i)}$ pour $k > 1$).

(d) On suppose que $m = 1$, donc que m est connu ($\lambda > 0$ restant inconnu). Quel est alors le modèle statistique ? Montrer que ce modèle appartient à la famille exponentielle, et en déduire une statistique exhaustive dont vous montrerez qu'elle est complète. Déterminer la matrice d'information de Fisher du modèle. Quelle est la fonction de λ (à une transformation affine près) qu'il est possible d'estimer efficacement ? Déterminer l'estimateur de maximum de vraisemblance de λ et en démontrer un théorème de la limite centrale.

(e) On suppose désormais que $m \in \mathbb{R}$ est inconnu, tout comme $\lambda > 0$. Quel est alors le modèle statistique ? Montrer que ce modèle n'appartient pas à la famille exponentielle. Allez à la question 2 de cet exercice, déterminer un estimateur (λ_n, m_n) du maximum de vraisemblance du couple (m, λ).

(f) Pour $a \in \mathbb{R}$, démontrer que $\frac{1}{n}\sum_{i=1}^{n} |X_i - a|$ converge presque sûrement quand $n \to +\infty$ vers $\mathbb{E}(|X - a|)$. Montrer que la fonction $a \mapsto \mathbb{E}(|X - a|)$ est minimale en $a = m$. En déduire que $m_n \xrightarrow[n \to +\infty]{p.s.} m$, puisque $\lambda_n \xrightarrow[n \to +\infty]{p.s.} \lambda$.

Université Paris Dauphine - Sorbonne

Première Année Master M.A.E.F. 2006 - 2007

Statistique

Examen de septembre 2007

Examen de 3 h 00. Tout document ou calculatrice est interdit.

1. Soit la fonction $f_a(x) = \frac{1}{2}(a - a^2 \cdot x) \cdot \mathbb{1}_{\{-1/a \le x \le 1/a\}}$ où $a > 0$.
   (a) Démontrer que f_a est une densité de probabilité par rapport à la mesure de Lebesgue et la tracer.
   (b) On suppose que X est une variable aléatoire de densité f_a. Déterminer $IE X$ et $\mathrm{var} X$.
   (c) Soit $(X_k)_{k \in \mathbb{N}^*}$ une suite de variables aléatoires indépendantes, telle que la densité de X_n soit f_n pour tout $n \in \mathbb{N}^*$. Déterminer la limite en probabilité de $(X_k)_{k \in \mathbb{N}^*}$ lorsque $n \to \infty$.
   (d) Soit $(X_1, \ldots, X_n)$ un échantillon de v.a.i.i.d. de densité f_a. On suppose que a est inconnu. Déterminer l'estimateur du maximum de vraisemblance a_n pour a. Calculer la fonction de répartition de a_n et en déduire sa convergence en probabilité vers a.
   (e) Pour $\alpha > 0$, déterminer un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour a.
   (f) Déterminer le test du rapport de vraisemblance de niveau α pour tester l'hypothèse $H_0 : a = a_0$, contre l'hypothèse $H_1 : a = a_0$ et déterminer la zone d'acceptation du test en fonction de α.
   (g) Proposer un autre estimateur convergent de a.

2. Une compagnie fabrique des piles et s'intéresse à savoir quelle est leur durée de vie moyenne T. Pour ce faire, on considère 10 000 piles produites le même jour que l'on soumet à la même contrainte. Comme on ne veut pas attendre que toutes les piles soient usées, on décide d'arrêter l'expérience au bout de 10 jours et de compter combien sont encore "en vie". Soit N_{10} ce nombre.
   (a) Dans une première approximation, on suppose que la durée de vie d'une pile peut être modélisée par une loi exponentielle de paramètre $\beta > 0$. Quelle est la loi de la durée de vie moyenne (théorique) T d'une pile en fonction de β ? Quelle est, en fonction de T, la probabilité qu'une pile meure avant 10 jours ? Montrer alors que $N_{10}/10\,000$ suit approximativement une loi dont on précisera les paramètres en fonction de T. En déduire alors un estimateur T de T en fonction de $N_{10}/10\,000$ dont on donnera une idée de la limite centrale.
   (b) Montrer que N_{10} n'est pas une statistique exhaustive pour T. Proposer un estimateur convergent de T, basé sur l'échantillon des 10 000 durées de vie des piles. Et si l'on avait attendu 12 jours au lieu de 10 ? Déterminer alors une région de confiance asymptotique pour T de niveau 95%.

mieux” T (donc trouver x tel que T soit de variance minimale). Ce résultat vous semble-t-il en pratique intéressant?